DOI:
10.1038/s41467-019-14204-z

Document Version
Peer reviewed version

A transcriptome-wide antitermination mechanism sustaining identity of embryonic stem cells

Yaroslav A. Kainov[1] and Eugene V. Makeyev[1,2]

1. Centre for Developmental Neurobiology, King's College London, London SE1 1UL, UK
2. Lead / Corresponding author:
Centre for Developmental Neurobiology
Guy's Hospital Campus, New Hunt's House
King's College London, London, SE1 1UL
United Kingdom
E-mail: eugene.makeyev@kcl.ac.uk

Summary

Eukaryotic gene expression relies on extensive crosstalk between transcription and RNA processing. Changes in this composite regulation network may provide an important means for shaping cell type-specific transcriptomes. Here we show that the RNA-associated protein Srrt/Ars2 sustains embryonic stem cell (ESC) identity by preventing premature termination of numerous transcripts at cryptic cleavage/polyadenylation sites in first introns. Srrt interacts with the nuclear cap-binding complex and facilitates recruitment of the spliceosome component U1 snRNP to cognate intronic positions. At least in some cases, U1 recruited in this manner inhibits downstream cleavage/polyadenylation events through a splicing-independent mechanism called telescripting. We further provide evidence that the naturally high expression of Srrt in ESCs offsets deleterious effects of retrotransposable sequences accumulating in its targets. Our work identifies Srrt as a molecular guardian of the pluripotent cell state.

Introduction

Eukaryotes are characterized by a remarkable degree of coordination between different steps of their gene expression program [1,2]. Most mRNA precursors (pre-mRNAs) are modified by the addition of a 7-methylguanosine cap to the 5' end, excision of introns by the spliceosome, and 3'-terminal cleavage and polyadenylation. Aberrant RNA species are degraded by specialized quality control mechanisms. All these events can occur co-transcriptionally, receiving regulatory inputs from elongating RNA polymerase II (pol II) but also modulating the efficiency of RNA synthesis through various forms of functional feedback [3-7].

Co-transcriptional capping of pol II transcripts followed by the assembly of the nuclear cap-binding complex (CBC) provides a critical line of communication between RNA synthesis and subsequent processing events [8,9]. The two core subunits of the CBC, Ncbp1/Cbc80 and the Ncbp2/Cbc20, can recruit several additional co-factors including the conserved multipurpose adapter protein Srrt/Ars2 [10-13]. Srrt has been shown to mediate degradation of promoter-proximal transcripts in an exosome-dependent manner, promote termination/3'-terminal maturation of replication-dependent histone mRNAs and several other pol II transcripts, and control production of small noncoding RNAs [10-12,14-16]. Of note, CBC can stimulate pre-mRNA splicing by recruiting U1 snRNP and other components of the spliceosome complex to cap-proximal introns [17-19], but whether this activity depends on Srrt is an open question.

Unlike the core CBC components expressed at relatively stable levels across different conditions, Srrt tends to be substantially more abundant in proliferating cells than in their differentiated or quiescent counterparts. Consistent with this behavior, Srrt has been shown to promote proliferation of mammalian cells both in vitro and in vivo [14,20,21]. These effects may be facilitated by the microRNA or/and histone mRNA regulation activities of Srrt [10,14,22,23]. On the other hand, Srrt contributes to maintenance of mouse neural stem cells (NSCs) in a

66 microRNA-independent manner, by promoting expression of the critical transcription factor
67 Sox2 [24]. Notably, Srrt is critical for early development in vertebrates [25,26]. However,
68 molecular mechanisms underlying this effect remain poorly understood.

69 Pre-mRNA cleavage and polyadenylation is another crucial point of gene regulation.
70 These two coupled reactions involve co-transcriptional assembly of multisubunit protein
71 complexes at a 6-nt polyadenylation signal (PAS) and its adjacent sequences, cleavage of the
72 nascent transcript at the cleavage/polyadenylation site (CS) located typically 10-30 nt
73 downstream of the PAS, and subsequent addition of a poly(A) tail to the newly formed 3' end
74 [27-29]. Co-transcriptional cleavage/polyadenylation triggers a rapid release of the elongating pol
75 II complex from the DNA template [30].

76 Interestingly, recruitment of U1 snRNP to 5' splice sites (5'ss) or other cognate motifs
77 can repress downstream CSs through a splicing-independent mechanism known as
78 telescripting [31,32]. Telescripting is required for normal expression of relatively long
79 mammalian genes [33], and its efficiency can be modulated by global changes in transcriptional
80 activity of the cell altering the ratio between free and pre-mRNA-associated U1 [32]. However,
81 it is unclear if telescripting can be controlled in a more nuanced cell type-specific manner.
82 Similarly, the emerging link between telescripting and early steps of pol II elongation awaits
83 further experimental characterization [34-36].

84 Embryonic stem cells (ESCs) are developmentally early progenitors capable of self-
85 renewal and differentiation into the three germ layers of the embryo proper. Several
86 transcription factors including Pou5f1/Oct4, Nanog and Sox2 are known to play a key part in
87 specifying molecular identity of this and other types of pluripotent stem cells [37-39]. Here we
88 identify Srrt as a top candidate in a screen for additional regulators involved in ESC
89 maintenance. We show that Srrt functions in this context by suppressing premature
90 termination of transcription at cryptic cleavage/polyadenylation sites in first introns. This

mechanism affects hundreds of genes active in ESCs and is mediated by CBC-dependent recruitment of U1 snRNP to 5'-proximal pre-mRNA sequences. In addition to its possible contribution to evolutionarily conserved gene regulation events, this activity limits deleterious effect of retrotransposable elements accumulating in first introns of its target genes. Overall, our work uncovers a transcriptome-wide antitermination circuitry with important roles in ESC biology.

Results

ESC maintenance depends on naturally high expression of Srrt

To understand possible role of RNA-based regulation mechanisms in maintenance of mouse ESCs, we inspected genes downregulated during neuronal and spontaneous differentiation of this cell type [40,41] (Fig. 1a). A stringent shortlisting procedure identified 84 top candidates with expression levels decreasing monotonically in both differentiation models (Supplementary Data 1). The list contained several previously characterized ESC-enriched transcription factors including but not limited to Pou5f1/Oct4 and Sox2 (Supplementary Data 1). Among putative regulators of RNA processing Srrt was a particularly promising candidate since its knockout (KO) results in preimplantation embryonic lethality [25] but its role in ESCs, i.e. cells matching this stage of mouse development, has not been investigated systematically.

Srrt protein was readily detectable in mouse ESCs and its levels were substantially reduced in proliferating NSCs [fold change (FC) 2.9; t-test p=1.3e-04] and post-mitotic neurons (FC=5.8; t-test p=8.8e-04; Fig. 1b). Srrt expression was also downregulated upon withdrawal of 2i inhibitors and LIF, the compounds required to maintain ESCs in an undifferentiated naïve state (Supplementary Fig. 1a, b; FC=2.4; t-test p=0.034; Ref[42]). Of note, the expression of the CBC subunit Ncbp1 remained constant under these conditions (Supplementary Fig. 1a, b; t-test p=0.78).

To address functional significance of the naturally high expression of Srrt in ESCs, we downregulated it to a level comparable to that observed in more differentiated cells using a mixture of four Srrt-specific siRNAs (siSrrt; Fig. 1c; compare with Fig. 1b and Supplementary Fig. 1a, b). This led to a loss of the characteristic rounded morphology of ESC colonies and reduced ESC-specific alkaline phosphatase activity compared to cultures treated with a control siRNA (siCtrl; Fig. 1d). Srrt knockdown also led to a readily detectable differentiation effect in a colony formation assay (Fig. 1e, f and Supplementary Fig. 1c-f). Moreover, siSrrt triggered a modest but statistically significant decrease in the expression of ESC-enriched surface markers SSEA1 and Pecam1/CD31 (Supplementary Fig. 1g, h). This suggests that maintenance of ESCs depends on relatively high expression of Srrt.

Srrt knockdown has a global effect on the ESC transcriptome

RNA sequencing (RNA-Seq) analysis uncovered considerable changes in the transcriptome of siSrrt-treated ESCs with 1828 downregulated and 1590 upregulated genes [FC≥1.5 and false discovery rate (FDR) <0.05; Supplementary Data 2]. The regulated genes showed a partial overlap with those changing their expression during spontaneous differentiation of ESCs (Supplementary Fig. 2a). Although expression of many pluripotency markers including Pou5f1/Oct4, Sox2 and Nanog remained unchanged in response to siSrrt, some examples of this category (e.g. Nr0b1, Pecam1 and Zic2) were detectably downregulated (Supplementary Fig. 2b). Conversely, expression of many developmental and differentiation markers increased (Supplementary Fig. 2b), in line with enrichment of corresponding gene ontology (GO) terms among the upregulated genes (Supplementary Data 3). For example, the GO terms developmental process, multicellular organismal development and cell differentiation were enriched with FDRs 3.6E-6, 7.4E-6 and 1.5E-5, respectively (Supplementary Data 3).

We confirmed RNA-seq expression data for 20 pluripotency and differentiation markers selected for RT-qPCR validation (Fig. 1g and Supplementary Fig. 1c).

Notably, downregulated genes were over-represented among the most reliable changes triggered by siSrrt (Supplementary Fig. 2d). Although we did not detect significantly enriched GO terms for this category of genes, some of the especially robust downregulation targets (FC≥2 and FDR<1E-50; dark red dots in Supplementary Fig. 2d) encoded known ESC markers and positive regulators of cell proliferation. Relevant examples included alkaline phosphatase Alpl (the enzyme assayed in Fig. 1d and Supplementary Fig. 1c-e), epigenetic regulator Cdyl2, activin receptor Acvr1b/Alk4, nuclear receptor co-activator Dcaf6/NRIP, and a conserved RAGNYA domain protein Ammecr1 mutated in the Alport syndrome with mental retardation, midface hypoplasia and elliptocytosis [43-47]. Downregulation of these genes was confirmed by RT-qPCR (Fig. 1g). Thus, Srrt may help ESCs to maintain their undifferentiated status by regulating extensive sets of genes.

Srrt limits expression of prematurely terminated transcripts

We noticed that many genes responded to Srrt knockdown by accumulating RNA-Seq reads in first (5'-proximal) introns (Supplementary Fig. 3a). This often coincided with downregulation of the corresponding genes (the lower right quadrant in Supplementary Fig. 3b and the blue line in Supplementary Fig. 3c) and when it did, the increase in the RNA-Seq coverage was strongly biased towards the 5' end of the first intron (Supplementary Fig. 3d). Relevant examples included the genes in the right plot in Fig. 1g (see below). To check if this behavior could be due to premature termination of transcription, we mapped the position of cleavage/polyadenylation sites (CSs) using 3'-proximal RNA-sequencing (3'RNA-Seq). This revealed a widespread activation of CSs within first introns in siSrrt-treated ESCs (Fig. 2a and Supplementary Fig. 4a).

Significant changes in premature cleavage/polyadenylation were less common in other introns and lacked the upregulation trend observed for first introns (Fig. 2a). Upregulated CSs in first introns tended to occur relatively close to the 5' splice site (5'ss) (Fig. 2b). Significantly fewer of these CSs were previously annotated in the polyA_DB3 database [48] compared to their counterparts located in 3'UTRs of the same genes (30.1% v 81.4%; Fisher's exact test p=3.9E-179). However, the incidence of canonical cleavage/polyadenylation signal (PAS) AATAAA or its common variant ATTAAA upstream of these two CS categories was virtually indistinguishable (Supplementary Fig. 4b). Hence, Srrt dampens the expression of multiple transcripts terminated at a poorly characterized class of CSs in first introns.

Srrt blocks cleavage/polyadenylation in first introns

Two possibilities could account for accumulation of prematurely terminated transcripts in response to Srrt knockdown: (1) enhanced pre-mRNA cleavage and polyadenylation at the corresponding intronic positions or (2) increased stability of these relatively short RNA species. The former mechanism should lower the production of full-length mRNA isoforms, while the latter is unlikely to produce this effect. Notably, activation of CSs in first introns strongly correlated with an overall decrease in expression levels of the corresponding genes (Fig. 2c, Supplementary Fig. 4c and Supplementary Data 4) and downregulation of CSs in their 3'UTRs (Supplementary Fig. 4d). There were 284 genes with intronic CS (iCS) upregulated ≥2-fold, FDR<0.05 and expression level reduced ≥1.5-fold, FDR<0.05, and an even larger number of genes showing this trend was detected using less stringent cutoffs (Supplementary Data 4). Genes upregulated despite the activation of iCSs were clearly a minority, and the increase in the overall expression levels in this case tended to be due to accumulation of prematurely terminated isoforms (e.g. the *Ttll11* gene in Supplementary Data 4).

RNA-Seq and 3'RNA-Seq coverage plots for individual targets were consistent with our transcriptome-wide analyses (Fig. 2d and Supplementary Fig. 5a). We used the 3'-terminal version of rapid amplification of cDNA ends (3'RACE) to map the regulated iCSs for three genes selected for experimental validation, *Ammecr1*, *Cdyl2* and *Dcaf6* (Supplementary Fig. 5b). In all three cases, siSrrt increased the RT-qPCR signal upstream of the iCSs and simultaneously reduced the abundance of downstream RNA sequences (Fig. 2e). This corresponded to a ~3-7-fold decrease in the ratio between the full-length and prematurely terminated transcripts, a statistic that we refer to as iCS readthrough efficiency (Supplementary Fig. 5c). A similar decrease in readthrough efficiency was evident when we substituted the siSrrt mixture with any of its 3 most efficient constituents, siSrrt#1, siSrrt#2 or siSrrt#3 (Supplementary Fig. 6a, b). The three individual siRNAs also caused largely similar to siSrrt effects on the expression of pluripotency and differentiation markers (Supplementary Fig. 6c-e).

To directly test the impact of intronic cleavage/polyadenylation on gene expression, we focused on *Ammecr1*. The overall expression of this biomedically important gene [45] decreased while the relative abundance of the iCS-terminated species increased during ESC differentiation into neurons, consistent with the *Srrt* downregulation trend (Supplementary Fig. 7a-d). Furthermore, knockdown of the full-length Ammecr1 transcripts induced detectable upregulation of a subset of the siSrrt-induced differentiation markers (Supplementary Fig. 7e, f). *Ammecr1* is encoded on the X chromosome, which also makes it an easy target for reverse genetics in male ESCs.

Importantly, when we deleted *Ammecr1* sequence containing two polyadenylation signals (PASs) upstream of the strongest Srrt-regulated iCS using CRISPR-Cas9 (Fig. 3a, b), the mutant allele (*ΔPAS*) lost its ability to undergo premature cleavage and reduce its expression output following Srrt knockdown (Fig. 3c-e). Together, these data suggest that Srrt

promotes expression of full-length mRNAs by blocking premature cleavage/polyadenylation in first introns.

iCS repression does not depend on the exosome or small RNAs

Since Srrt has been previously shown to destabilize transcription start site (TSS)-proximal transcripts in an exosome-dependent manner [12], we compared our 3'RNA-Seq data with results of 3' end-proximal RNA-Seq (2P-Seq) for mouse ESCs where the exosome complex was inactivated by knockout of its core subunit Exosc3 [36]. Metaplot analysis of siSrrt-regulated genes showed a robust accumulation of TSS-proximal RNAs transcribed in the sense but not the antisense direction (Supplementary Fig. 8a). On the other hand, Exosc3 KO increased the abundance of both sense and antisense transcripts in the same genomic regions (Supplementary Fig. 8b), as described previously [36].

In stark contrast to siSrrt, Exosc3 KO had no detectable effect on the abundance of full-length mRNAs transcribed from Srrt-dependent genes (Supplementary Fig. 8c). Although downregulation of the catalytic exosome subunits Exosc10 and Dis3 by corresponding siRNAs promoted some accumulation of prematurely terminated Ammecr1 RNA (Supplementary Fig. 8d, e), neither these nor an Exosc3-specific siRNA decreased the abundance of full-length Ammecr1 transcripts (Supplementary Fig. 8d, e). Conversely, exosome-specific siRNAs caused more efficient accumulation of TSS-proximal upstream antisense transcripts compared to siSrrt (Supplementary Fig. 8e).

To check the possibility that intronic cleavage/polyadenylation might be controlled through Srrt-stimulated production of small noncoding RNAs [10,14,16], we turned to published RNA-seq data for Dicer1/Dicer KO in mouse ESCs with a validated effect on microRNA activity [49]. The gene expression changes induced by Srrt knockdown and Dicer1 KO showed no global correlation (Supplementary Fig. 9a) and the expression of Srrt-regulated genes did

not generally change in response to Dicer1 KO (Supplementary Fig. 9b). Moreover, inspection of RNA-seq coverage profiles for individual Srrt targets showed no evidence for iCS regulation by Dicer (Supplementary Fig. 9c). Thus, neither the exosome nor small RNAs appear to be required for Srrt-mediated repression of intronic cleavage/polyadenylation in mouse ESCs.

Srrt-mediated repression of iCSs relies on the CBC

To examine possible contribution of the CBC to the Srrt-dependent antitermination activity, we knocked down Ncbp1 in mouse ESCs and compared the effect of this treatment with that induced by siSrrt (Fig. 4a). RNA-Seq and 3'RNA-Seq analyses revealed a noticeable correlation between the siNcbp1- and the siSrrt-treated samples in terms of overall gene expression changes and activation of CSs in first introns (Fig. 4b, c and Supplementary Fig. 10a-c).

To test if Srrt and Ncbp1 functioned in the same pathway, we generated an ESC line containing a doxycycline (Dox) inducible human SRRT transgene (SRRT-Tg) resistant to mouse-specific siSrrt (Fig. 4d and Supplementary Fig. 10d). Importantly, SRRT-Tg was sufficient to rescue termination of Ammecr1 transcripts in the first intron induced by siSrrt but not by siNcbp1 (Fig. 4e, f). In line with this functional interaction between the two proteins and published data for their human counterparts [11,12], Srrt and Ncbp1 interacted physically in mouse ESCs in a nucleic acid-independent manner (Supplementary Fig. 10e). RNA immunoprecipitation (RIP) with Ncbp1-specific antibodies showed that siSrrt did not alter the ability of Ncbp1 to interact with (pre-)mRNAs (Supplementary Fig. 10f), suggesting that Ncbp1 might be required for recruiting Srrt to its targets but not the other way around.

We concluded that the ability of Srrt to repress cleavage/polyadenylation in first introns depends on its interaction with the CBC.

Srrt facilitates U1 binding upstream of regulated iCSs

CBC can promote recruitment of U1 to cap-proximal introns, and this snRNP can in turn antagonize cleavage/polyadenylation via telescripting [18,31]. To assess possible contribution of these mechanisms, we mapped U1-binding sites in formaldehyde-crosslinked ESCs using RNA antisense purification-sequencing (RAP-Seq; Ref[50]; Fig. 5a). We ascertained that the U1 pull-down procedure worked successfully by monitoring enrichment of U1 snRNA precursors and depletion of the 45S ribosomal RNA (Supplementary Fig. 11a). Reflecting the known U1 interaction preferences, input-normalized RAP-Seq reads showed a detectable bias towards the 5' end of all introns and first introns containing Srrt-repressed iCSs (Supplementary Fig. 11b, c).

Although the siCtrl- and the siSrrt-treated ESCs showed generally similar U1 binding profiles (Supplementary Fig. 11b, c), we noticed a discernable U1 peak upstream of the Srrt-regulated iCSs in the siCtrl but not the siSrrt sample (Supplementary Fig. 11d). Supporting this observation, the incidence of U1 clusters deduced using a previously described approach [51] was significantly higher in a 250-nt window upstream of Srrt-repressed iCSs than in a similarly sized downstream window in the siCtrl-treated cells (Fig. 5b). This was consistent with enrichment of relatively strong U1 binding motifs upstream of iCSs compared to corresponding downstream positions and 250-nt windows adjoining CSs in 3'UTRs of the same genes (Fig. 5c). Importantly, Srrt knockdown led to a significant drop in U1 cluster coverage upstream of the regulated iCSs (Fig. 5b).

The above effects were also detectable for individual Srrt targets. For example, two prominent U1 RAP-Seq peaks between the 5'ss and the strongest Srrt-repressed CSs in the first intron of the *Ammecr1* gene were significantly enriched over the input in the siCtrl- but not the siSrrt-treated samples (Fig. 5d). RT-qPCR analyses of the pull-down and the input

fractions confirmed that U1 binding to the corresponding intronic positions was significantly reduced by Srrt knockdown (Fig. 5e). In contrast, U1 occupancy in the first intron of *Ncbp2*, a control gene not regulated by Srrt, showed no significant difference between the siCtrl and siSrrt samples (Fig. 5e and Supplementary Fig. 11e).

The siSrrt effect on U1 recruitment was not due to major changes in U1 snRNA steady-state levels or its processing efficiency (Supplementary Fig. 12a, b). The levels of the U1 snRNP proteins Snrpa/U1-A and Snrp70/U1-70K were also unaffected (Supplementary Fig. 12c, d). Furthermore, we compared our 3'RNA-seq data for siSrrt-treated samples with a similar analysis published for mouse ESCs where U1 was inactivated by an antisense morpholino oligonucleotide [36]. Although both treatments promoted premature cleavage/polyadenylation in first introns, inactivation of U1 clearly differed from Srrt knockdown by additionally inducing this effect in non-first introns on a transcriptome-wide scale (Supplementary Fig. 12e, f).

These data suggest that Srrt facilitates U1 recruitment upstream of regulated CSs in first introns rather than substantially altering overall activity of this snRNP in mouse ESCs.

Srrt-recruited U1 can promote telescripting

As a direct test of the U1 effect on iCSs, we treated ESCs with a U1-specific antisense morpholino oligonucleotide (amoU1; Fig. 6a). This enhanced the efficiency of premature cleavage-polyadenylation in the first intron of Ammecr1 pre-mRNA compared to samples treated with a non-targeting control (amoCtrl) or an antisense morpholino against another spliceosomal snRNA, U2 (amoU2). The noticeably stronger effect of amoU1 than that of amoU2 suggested that Srrt-stimulated recruitment of U1 snRNP could inhibit iCSs through telescripting rather than the spliceosome assembly pathway.

To test this hypothesis, we prepared a minigene construct by fusing the exon 1-intron 1 junction and the Srrt-regulated iCS region of the *Ammecr1* gene with a recombinant 3'UTR containing a constitutive CS (Fig. 6b). Since it lacked a functional 3'ss, this cassette allowed us to assay telescripting in the absence of pre-mRNA splicing. The minigene was expressed in ESCs pretreated with siSrrt or siControl, and the use of the Ammecr1 iCS was analyzed by RT-qPCR (Fig. 6c). Recapitulating the behavior of endogenous Ammecr1 pre-mRNAs, minigene-derived transcripts showed more efficient iCS readthrough in the siCtrl than in the siSrrt samples (Fig. 6c).

Mutation of the 5'ss, i.e. the site where U1 binds to initiate splicing of endogenous Ammecr1 transcripts, had no detectable effect on the minigene response to siSrrt (Fig. 6c). However, when we mutated three additional positions predicted to interact with U1, the minigene was terminated at the iCS regardless of the Srrt expression levels (Fig. 6c). On the other hand, deletion of the PAS hexamers (ΔPAS) preceding the iCS led to a constitutive readthrough phenotype (Fig. 6c).

These results confirm that Srrt can block intronic cleavage/polyadenylation through a U1-dependent telescripting mechanism.

Many iCSs emerged through retrotransposition

Our data so far suggested that productive transcription of a large subset of genes active in ESCs depends on Srrt abundance. To understand evolutionary mechanisms underlying this regulation, we examined interspecies conservation scores [52] for 50-nt windows bounded by 40 nt upstream and 10 nt downstream of Srrt-regulated iCSs (Fig. 7a). A fraction of these sequences (39.6%) showed detectable conservation (average PhastCons score >=0.1). This category included *Ammecr1*, *Cdyl2* and *Dcaf6*, which had their iCS-associated PAS hexamers present in several mammalian species (Supplementary Fig. 13).

A majority of the Srrt-regulated sequences (60.4%) were conserved poorly or not at all (average PhastCons score <0.1). Since retrotransposable elements (RTEs) provide an important source of interspecies diversity [53,54], we wondered if mouse/rodent-specific iCSs could appear as a result of relatively recent retrotransposition events. Strikingly, an RTE density plot revealed a prominent peak of these elements integrated in the sense orientation immediately upstream of the Srrt-repressed iCSs (Fig. 7b). Conversely, antisense RTE sequences were depleted in this region (Fig. 7b).

The iCS-associated sense-strand peak was ~200-nt wide suggesting that it could be dominated by relatively short RTEs (Fig. 7b). Indeed, most of the sense-strand RTEs that terminated around an iCS (±50 nt) belonged to the group of short interspersed nuclear elements (SINEs), although a few long interspersed nuclear elements (LINEs) and long terminal repeats (LTRs) were also detected (Fig. 7c) [53,54]. Members of the B2 SINE family were especially common at this position (Supplementary Fig. 14a), consistent with the presence of canonical PASs in their consensus sequence [55]. Overall, 31.2% of all regulated iCSs were associated with 3' ends of sense-strand RTEs.

iCS-associated B2 SINEs were found for example in genes encoding activin receptor Acvr1b (see also Fig. 1g), WNT pathway modulator Ankrd6/Diversin, Down Syndrome critical region protein Dscr3, and heat shock protein-associated factor Hspbap1 (Fig. 7d and Supplementary Data 4; https://www.genecards.org). Genes with iCSs occurring at the end of a LINE repeat included those encoding ankyrin repeat and SOCS box protein Asb3 and a component of a regulatory complex interacting with unmethylated DNA in ESCs, Zbtb25 (Fig. 7e and Supplementary Data 4; https://www.genecards.org). In many cases, PAS hexamers preceding iCSs matched corresponding elements in the parental RTEs (Fig. 7d, e).

iCSs occurring at the 3' end of sense-strand RTEs were significantly less conserved than the rest of the iCSs (Fig. 8a), suggesting that the corresponding RTE sequences might be

a result of relatively recent jumps. Indeed, the iCS-associated repeats were less divergent from the master copies, as compared to control groups comprising all sense or antisense repeats from first introns or the entire collection of repeats found in the mouse genome (Fig. 8b).

Regardless of the RTE association status of their iCSs, all Srrt-regulated first introns showed a significantly higher density of RTE-derived sequences compared to non-regulated first or non-first introns (Fig. 8c and Supplementary Fig. 14b). We also observed a strong bias towards antisense orientation of RTEs in all groups of introns (Fig. 8c), suggesting that sense-oriented RTEs might be more disruptive and therefore subject to stronger purifying selection than their antisense counterparts.

We concluded that, in addition to controlling evolutionarily conserved events, Srrt might repress deleterious iCSs appearing as a result of retrotransposition.

Srrt target genes tend to have long RTE-rich first introns

Telescripting is known to be critical for production of long transcripts [33]. Interestingly, we detected a genome-wide correlation between the RTE density and the overall size of first introns (Fig. 8d). In line with their increased RTE load, first introns of Srrt-dependent genes tended to be significantly longer compared to control groups (Fig. 8e). Of note, Srrt-regulated and non-regulated first introns were indistinguishable based on their 5'ss strength (Supplementary Fig. 14c).

To find out if the length of first introns might be a good predictor of the Srrt dependence, we plotted average rpkm values in control-treated ESCs for genes separated into three equally sized groups according to the length of their first intron (short, mid and long; Fig. 8f). Genes with longer first introns tended to be expressed at lower levels in ESCs even in the presence of normal amounts of Srrt. The presence of one or more AATAAA hexamers

in the first intron was associated with somewhat reduced average expression in each category, but this effect was not statistically significant (Fig. 8f). Notably, the length of the first intron showed a strong positive association with the ability of AATAAA to dampen gene expression in response to Srrt knockdown (Fig. 8g).

Thus, recurrent RTE jumps may sharpen the dependence of gene expression on Srrt by increasing the length of first introns.

Discussion

Our study uncovers a global antitermination mechanism responsible for productive expression of multiple genes in pluripotent stem cells (Fig. 8h). This mechanism relies on the ability of Srrt to associate with the CBC and block premature cleavage/polyadenylation of pre-mRNAs in first introns by promoting recruitment of U1 snRNP to cap-proximal sequences. We show that, at least in the case of the disease-associated gene *Ammecr1*, Srrt-augmented U1 binding can promote transcriptional readthrough of a downstream iCS as a result of telescripting.

Three lines of evidence argue that Srrt is an important regulator of ESC identity. (1) Srrt is substantially more abundant in ESCs than in other cell types including actively proliferating NSCs (Fig. 1b and Supplementary Fig. 1a, b). (2) Normal expression of hundreds of iCS-containing genes active in ESCs relies on the naturally high levels of Srrt (Fig. 2c, Supplementary Fig. 4d and Supplementary Data 4). (3) Srrt downregulation in ESCs to levels considered physiological in other cell types induces several differentiation-specific changes (Fig. 1b-g and supplementary Figs. 1 and 2a-c). It is possible that the latter effect depends, at least in part, on reduced expression of a subset of the iCS genes. Indeed, knockdown of *Ammecr1* leads to statistically significant upregulation of some differentiation markers induced in response to Srrt-specific siRNAs (Supplementary Fig. 7f). Further

research will be required to understand molecular functions of the Ammecr1 protein and identify other Srrt targets that may contribute to the ESC differentiation phenotype.

The role of Srrt in ESCs appears to be distinct from its function as a transcriptional activator of *Sox2* gene in NSCs [24]. Sox2 mRNA levels did not change in our siSrrt-treated samples implying that other mechanisms must ensure robust expression of this important transcription factor in ESCs. This may be achieved through cross-activation of *Sox2* by Pou5f1, Nanog or other transcriptional regulators present in ESCs but not NSCs [37-39]. Alternatively, it is possible that the residual amount of Srrt protein in siSrrt-treated ESCs (Fig. 1c) is sufficient for promoting *Sox2* transcription but not for blocking iCSs. Consistent with a possible difference in quantitative requirements of the two mechanisms, Srrt is ~3 times more abundant in ESCs than in NSCs cultured in vitro (Fig. 1b).

Our data support the emerging view that, in addition to their reliance on transcription factors, pluripotent stem cells depend on adequate expression patterns of a number of RNA-associated proteins. These include for example pre-mRNA splicing regulators identified in recent studies [56-59]. It is likely that further quantitative analyses of expression changes triggered by ESC differentiation or transition of differentiated cells to induced pluripotency will uncover additional factors altering RNA processing and tuning the way it communicates with transcription.

Mounting evidence suggests that U1 snRNP-dependent readthrough of premature CSs is a widespread mechanism facilitating efficient transcription of long mammalian genes [31,33]. Furthermore, many pol II promoters are inherently bidirectional and the preferred direction for productive elongation appears to be selected based on the ability of promoter-proximal RNA sequences to recruit U1 snRNPs and limit the effect of premature cleavage/polyadenylation [34-36]. Interestingly, the efficiency of telescripting can be modulated by dynamic interactions between the U1 snRNP and nascent pre-mRNA pools, linking rapid

transcriptional activation in cells responding to external cues with corresponding changes in alternative cleavage/polyadenylation patterns [32].

We extend this line of research by showing that the ability of U1 to inhibit cryptic CSs can be tuned depending on the cell type and the 5' to 3' position of regulated sequences. This regulation logic is conceptually similar to prokaryotic antitermination used for example by bacteriophage λ to switch between immediate and delayed early stages of its gene expression program [60]. Despite fundamental mechanistic differences both systems rely on elevated expression of key RNA-associated factors, Srrt in ESCs and the N protein in λ, to repress transcription termination signals.

We cannot currently rule out that, in a subset of genes, Srrt-recruited U1 may antagonize intronic cleavage/polyadenylation through kinetic competition with splicing, instead of or in addition to telescripting. Supporting possible involvement of Srrt in splicing, some of its targets not regulated at the level of mRNA abundance appear to retain first introns in siSrrt-treated ESCs (yellow line in Supplementary Fig. 3d). Moreover, Srrt is known to control splicing decisions in plants [61,62]. What might determine the choice between telescripting- and splicing-dependent mechanisms on a transcriptome-wide scale is an interesting question for future studies.

It will be also important to understand how different molecular activities of Srrt are balanced depending on the cell type and RNA target identity. Especially intriguing is the ability of Srrt to promote 3'-terminal processing/termination in some cases [11,12,14,63] while antagonizing it in a transcriptome-wide manner in mouse ESCs (Fig. 2c, Supplementary Fig. 4d and Supplementary Data 4). We envisage at least two non-mutually exclusive explanations. (1) Srrt may block cleavage/polyadenylation only in the presence of sufficiently strong U1 binding motifs between the 5'-terminal cap and the iCS. In addition to promoting telescripting, U1 recruited to these positions might potentially compete with

cleavage/polyadenylation machinery for overlapping interaction sites in the Srrt protein. (2) Alternatively, ESCs may express yet-to-be identified Srrt-associated factors overriding the ability of this multipurpose adaptor to stimulate cleavage/polyadenylation or/and strengthening its contacts with U1.

Several Srrt-regulated iCSs appear to be conserved in evolution (Fig. 7a and Supplementary Fig. 13), pointing at their potential adaptive value. For example, such intronic elements may limit the abundance of ESC-enriched transcripts in other cell types. Supporting this possibility, the progressive decline in *Ammecr1* expression during neuronal differentiation correlates positively with the Srrt downregulation trend and negatively with an increase in the relative abundance of iCS-terminated Ammecr1 transcripts (Supplementary Fig. 7a-d). However, most iCSs lack detectable interspecies conservation and many of them are associated with relatively recent retrotransposition events (Fig. 7 and Fig. 8a, b).

What could be the role of Srrt in this context? Interestingly, Srrt-regulated first introns have a higher RTE load compared to non-regulated first and non-first introns (Fig. 8c and Supplementary Fig. 14b). This might reflect possible integration bias of RTEs to open chromatin, making first introns in genes transcriptionally active at the preimplantation stage especially vulnerable to recurrent and potentially heritable retrotransposition [64-66]. Accumulation of RTEs in this region would in turn dampen gene expression by introducing PASs/iCSs directly (Fig. 7) or making the acquisition of new PAS-like mutations more likely due to an increase in intron length (Fig. 8c-g and Supplementary Fig. 14b).

We propose that the natural over-expression of Srrt helps ESCs to alleviate potentially damaging consequences of this genome-wide effect. The largely negative impact of RTEs on individual fitness is often discussed in conjunction with their role as an important source of evolutionary innovation [53,54,67-70]. Hence, an intriguing possibility that should be investigated in the future is that, besides protecting the transcriptome, Srrt may also function as a genetic

capacitor allowing initially deleterious events to be repurposed for building new regulation modules.

Methods

Cell culture techniques

A2lox mouse ESCs [71] were cultured in a humidified incubator at 37°C, 5% CO_2, in plates or dishes coated with gelatin (Millipore, cat# ES-006-B) in 2i medium [37] containing a 1:1 mixture of Neurobasal (Thermo Fisher Scientific, cat# 21103049) and DMEM/F12 (Sigma, cat# D6421) media supplemented with 100 units/ml PenStrep (Thermo Fisher Scientific, cat# 15140122), 1 μM PD03259010 (Cambridge Bioscience, cat# SM26-2), 3 μM CHIR99021 (Cambridge Bioscience, cat# SM13-1), 0.5 mM L-Glutamine (Thermo Fisher Scientific, cat# 25030024), 0.1 mM β-mercaptoethanol (Sigma, cat# M3148), 1,000 units/ml ESGRO LIF (Millipore, cat# ESG1107), 0.5× B-27 supplement without vitamin A (Thermo Fisher Scientific, cat# 12587010) and 0.5× N2 supplement. N2 100× stock was prepared using DMEM/F12 medium as a base and contained 5 mg/ml BSA (Thermo Fisher Scientific, 15260037), 2 μg/ml progesterone (Sigma, P8783-1G), 1.6 mg/ml putrescine (Sigma, P5780-5G), 3 μM sodium selenite solution (Sigma, S5261-10G), 10 mg/ml apo-transferrin (Sigma, T1147-100MG), and 1 mg/ml insulin (Sigma, I0516-5ML) and stored in single-use aliquots at -80°C.

Cells were typically passaged every 2-3 days by treating the cultures with 0.05% Trypsin-EDTA (Thermo Fisher Scientific, cat#15400054) for 8-10 min at 37°C. After quenching trypsin with FBS (Thermo Fisher Scientific, cat# SH30070.03E), cells were washed once with Neurobasal medium and plated at a 1:6 dilution.

For RNA interference (RNAi) experiments, 2×10^5 cells were seeded in 1 ml of 2i medium per gelatinized well of a 12-well and immediately transfected with 50 pmol of an appropriate siRNA (Horizon Discovery; see Supplementary Data 5 for details) premixed with 3 μl of Lipofectamine 2000 (Thermo Fisher Scientific, cat# 11668019) and 100 μl of Opti-MEM I (Thermo Fisher Scientific, cat# 31985070), as recommended. The cultures were then incubated for 48 hours without changing the medium. In minigene experiments, cells pre-treated with siRNAs for 24 hours were transfected with 500 ng of minigene plasmid mixed with 2 μl of Lipofectamine 2000 and 100 μl of Opti-MEM I and incubated for another 24 hours prior to RNA extraction.

Stable knock-in lines were generated as follows. A2lox cells were pre-treated overnight with 1 μg/ml doxycycline (Dox; Sigma, cat# D9891-1G) to activate Cre expression, trypsinized and then transfected in suspension with 1 μg of an appropriate p2Lox-based plasmid mixed with 3 μl of Lipofectamine 2000 and 100 μl of Opti-MEM I in 4 ml of 2i medium in 6 cm bacterial dishes at $0.75\text{-}1\times10^5$ cells/ml. Cells were collected 2 hours post transfection and serially diluted in 2i medium prior to re-plating in 6-well format. On the next day, 350 μg/ml of geneticin/G418 (Sigma, cat# 10131019) was added and the incubation was continued for an additional 8-12 days with regular medium changes to allow geneticin-resistant cells to form colonies. These were picked, expanded and analyzed for inducible expression of transgenic sequences using reverse transcriptase-quantitative PCR (RT-qPCR) and/or immunoblotting.

Genomic deletions were generated in A2Lox cells containing a Dox-inducible Cas9 transgene. Cells were pre-treated with 1 μg/ml Dox overnight, transfected with a mixture containing two synthetic EditR gRNAs flanking the deletion region (50 pmol each; Horizon

Discovery; see Supplementary Data 5) or two EditR Non-targeting control gRNAs (50 pmol each; Horizon Discovery, cat# U-007501-01-05 and U-007501-01-05) and 100 pmol of synthetic EditR tracrRNA (Horizon Discovery, cat# U-002005-05) at 1-2×10^5 cells per well of 12 well plate using conditions described for RNAi experiments. Cells were trypsinized 24 hours post transfection, FBS-quenched, passed through Falcon 40 μm cell strainers (Corning, cat# 352340) to obtain a single-cell suspension, and serially diluted in 2i medium prior to re-plating in 6-well format. The cultures were then maintained for 8-12 days with regular medium changes and colonies originating from individual cells were picked, expanded, and their genomic DNA was analyzed for the presence of desired deletion using PCR genotyping (see below).

For antisense morpholino oligonucleotide (AMO) delivery, 2×10^6 ESCs were electroporated in the presence 7.5 μM of U1-specific, U2-specific or a scrambled AMO (Gene Tools, LLC; see Supplementary Data 5) in Amaxa Nucleofector II (Lonza) using ESC-specific program A-23 and Mouse Embryonic Stem Cell Nucleofector Kit (Lonza, cat# VPH-1001) as recommended. Nucleofected cells were maintained in 2i medium in a single well of a 6-well plate for 8 hours prior to RNA purification and RT-qPCR analysis.

Pluripotency/differentiation assays

To assess gene knockdown effects on ESC pluripotency/differentiation status, siRNA-transfected cells were incubated in 2i medium supplemented with 2% FBS for 48 hours and stained using an alkaline phosphatase detection kit (Millipore, cat# SCR004) as recommended. In colony formation assays, siRNA-transfected cells were trypsinized 24 hours post transfection, quenched with FBS, passed through Falcon 40 μm cell strainers, and plated at 1000 cells per well of a 6-well plate in 2i medium supplemented with 2% FBS. Seven days post plating cell colonies were stained for alkaline phosphatase, imaged, and analyzed using ImageJ (https://imagej.nih.gov/ij/; see Supplementary Data 5 for further information on the computer software used in this study).

For flow cytometry, ESCs transfected with siRNAs in a 12-well plate format were incubated in 2i medium for 48 hours, dissociated using Accutase (Thermo Fisher Scientific, cat# A1110501), washed with 1×PBS, pH 7.4 (Thermo Fischer Scientific, cat# 10010023), and resuspended in 100 μl of FACS buffer containing 1×PBS, 2 mM EDTA and 3% FBS. Cells were then stained for ESC surface markers using an APC-conjugated anti-Pecam1/CD31 antibody (Thermo Fisher Scientific, cat# 17-0311-80, 0.5μg per test) and an Alexa Fluor 488-conjugated anti-SSEA1 antibody (Thermo Fisher Scientific, cat# 53-8813-41, 0.125μg per test) for 1 hour on ice, washed twice with 300 μl of the FACS buffer and passed through Falcon 40 μm cell strainers to obtain single-cell suspensions. Samples were supplemented with 0.2 μg/ml DAPI ~10 min prior to flow cytometry to label membrane-compromised cells. Cells were then analyzed using a BD FACSCanto™ II cytometer equipped with 405 nm, 488nm and 633 nm lasers. The FCS files were analyzed using the flowCore and the flowViz packages (https://www.bioconductor.org/packages/release/bioc/html/flowCore.html; https://www.bioconductor.org/packages/release/bioc/html/flowViz.html). The following gating strategy was applied to select individual living (DAPI-negative) cells:

```
rg <- rectangleGate(filterId="myFilter", "FSC.A"=c(60000, 140000),
"SSC.A"=c(20000, 130000), "SSC.W"=c(80000, 160000), "DAPI.A" = c(-100, 5000))
```

The Pecam1 (APC) and SSEA1 (Alexa Fluor 488) signals were then measured in cells passing these gates (>28,000 per sample).

DNA constructs

Plasmids p2lox and pX330-U6-Chimeric_BB-CBh-hSpCas9 were kindly provided by Michael Kyba (Addgene plasmid #34635; Ref[71]) and Feng Zhang (Addgene plasmid #42230; Ref[72]). pEGFP-N3 was from Clontech and the pCR-bluntII-topo clone containing full-length open reading frame of human *SRRT* was from Horizon Discovery (MGC Human SRRT Sequence-Verified cDNA, Accession: BC109117, Clone ID: 40035609 cat# MHS6278-211690300). New constructs were generated as described in Supplementary Data 6 using routine molecular cloning techniques and enzymes from New England Biolabs. *Ammecr1* minigene plasmids were mutagenized as outlined in Supplementary Data 6 using a modified Quikchange site-directed mutagenesis protocol, in which PfuTurbo was substituted with the KAPA HiFi DNA polymerase (Kapa Biosystems, cat# KK2101). All constructs were verified by Sanger sequencing. Maps of all constructs are available on request.

PCR genotyping

Genomic DNA was prepared and analyzed using PCRBIO Rapid Extract PCR Kit (PCR Biosystems; cat# PB10.24-08) according to manufacturer's protocol. Amplified DNA fragments were resolved by electrophoresis in 1-2% agarose gels alongside GeneRuler 1 kb Plus DNA Ladder (Thermo Fisher Scientific, cat# SM1331). Deletion of a cleavage/polyadenylation site-containing fragment in the *Ammecr1* gene was confirmed using Ammecr1_genotype_F/Ammecr1_genotype_R primers (Supplementary Data 7) and Sanger sequencing of the PCR product.

RNA purification and RT-qPCR analyses

Total RNAs for gene expression analyses were extracted using an EZ-10 DNAaway RNA Miniprep Kit (BioBasic, cat# BS88136). Reverse transcription (RT) was performed at 50°C for 30 min using SuperScript IV reagents (Thermo Fisher Scientific, cat# 18090200) supplemented with 5 µM of random decamer (N10) primers and 2 units/µl of murine RNase inhibitor (New England Biolabs, M0314L). cDNA samples were analyzed by qPCR using a Light Cycler®96 Real-Time PCR System (Roche) and qPCR BIO SyGreen Master Mix (PCR Biosystems; cat# PB20.16). In minigene experiments, total RNAs were isolated from cells using TRIzol (Thermo Fisher Scientific, cat# 15596026), as recommended, with an additional acidic phenol-chloroform (1:1) extraction step. The aqueous phase was precipitated with an equal volume of isopropanol, washed with 70% ethanol and rehydrated in 80 µl of nuclease-free water (Thermo Fisher Scientific, cat# AM9939). RNA samples were then treated with 4-6 units of Turbo DNase (Thermo Fisher Scientific, cat# AM2238) at 37°C for 30 min to remove the bulk of DNA contaminants, extracted with equal volume of acidic phenol-chloroform (1:1), precipitated with 3 volumes of 100% ethanol and 0.1 volume of 3 M sodium acetate (pH 5.2), washed with 70% ethanol and rehydrated in nuclease-free water. Remaining traces of DNA were removed by pre-treating RNA samples with 2 units of RQ1-DNAse (Promega, cat# M6101) per 1 µg of RNA at 37°C for 30 min. RQ1-DNAse was inactivated by adding the stop solution as recommended and the RNAs were immediately reverse-transcribed using SuperScript IV and random decamer (N10) primers at 50°C for 30 min. All RT-qPCR primers are listed in Supplementary Data 7. Unless mentioned otherwise, RT-qPCR signals were normalized to expression levels of the Cnot4 housekeeping mRNA. In RAP and RIP RT-qPCR assays, signals in pull-down fractions were normalized to input signals obtained using the same primer pair. In minigene experiments, the RT-qPCR signals detected using primers annealing downstream of the Ammecr1 iCS were normalized to those obtained using upstream primers (see Supplementary Data 7 and Fig. 6b).

3'RACE

3'RACE was performed in principle as described [73]. Briefly, total RNAs were extracted from siSrrt-transfected ESCs using an EZ-10 DNAaway RNA miniprep kit. The RT step was done at 50°C for 60 min using SuperScript IV reagents, 5 μM of the 3'RACE_RT primer (Supplementary Data 7) and 2 units/μl of murine RNase inhibitor. This was followed by two rounds of nested PCR using PCRBIO Ultra Mix Red (PCR Biosystems, PB10.33-05): (1) with the 3'RACE_Q0 primer and a gene-specific primer GS1 and (2) with 3'RACE_Q1 primer and a gene-specific primer GS2 (Supplementary Data 7). The PCR products were then agarose gel-purified using a NucleoSpin gel and PCR clean-up kit (Macherey Nagel cat# 740609.250) and analyzed by Sanger sequencing.

Northern blotting

Northern blotting was performed using a DIG Northern starter kit (Merck, cat# 12039672910), as recommended. To prepare a U1-specific antisense digoxigenin-labeled probe, pML475 plasmid (Supplementary Data 6) was linearized with PvuII (New England Biolabs), purified using a NucleoSpin gel and PCR clean-up kit, and used as a template for SP6 RNA polymerase. 2.0×10^6 A2lox ESCs were plated in 10 cm gelatinized cell culture dishes in 10 ml of 2i medium and immediately transfected with 500 pmol of either siCtrl or siSrrt premixed with 27 μl of Lipofectamine 2000 and 1.5 ml of Opti-MEM I. Total RNAs were extracted 48 hours post transfection using TRIzol as described above. Purified RNA samples were dissolved in nuclease-free water at ~1 μg/μl and 2-μg aliquots were mixed with 8 μl of the gel loading buffer containing 98% Formamide (Thermo Fisher Scientific, cat# 15515026), 10 mM EDTA, 200 μg/ml bromophenol blue (Thermo Fisher Scientific, cat# 10243420), and 200 μg/ml xylene cyanol (Severn Biotech Ltd, cat# 30-60-01). The samples were then denatured at 70°C for 3 min, chilled on ice, and resolved by electrophoresis in 8% polyacrylamide gels (acrylamide:bis 29:1; Severn Biotech Ltd, cat# 20-3500-05) containing 8 M urea (Thermo Fischer Scientific, cat# 15505-027) and 1×TBE (Sigma, cat# T4415). RNAs were transferred from the gels to Hybond-N+ membranes (Merck, cat# GERPN1210B) using a Trans-Blot SD semi-dry transfer cell (Bio-Rad) in 0.5×TBE at 3 mA/cm^2. Membrane were stained with 0.02% methylene blue (Fisher Scientific, cat# 11443697) in 0.3 M sodium acetate pH 5.2 (Sigma, cat# S7899) and photographed. After destaining in 0.2×SSC (Sigma, cat# S6639) and 1% SDS (Promega, cat# H5114) membranes were blocked with DIG Easy Hyb solution at 68°C for 30 min and hybridized with 100 ng/ml probe in DIG Easy Hyb solution at 68°C overnight. Membranes were then washed twice in 2×SSC with 0.1% SDS at room temperature and twice in 0.1×SSC with 0.1% SDS at 68°C, 5 min each wash. The subsequent steps were done at room temperature. Membranes were washed in the Washing buffer containing 0.1 M maleic acid-NaOH, pH 7.5 (Sigma, cat# M0375), 0.15 M NaCl (Sigma, cat# 71376-1KG) and 0.3% (v/v) Tween 20 (Sigma, cat# P9416) for 5 min and blocked in 1× DIG Northern starter kit blocking solution for 30 min. This was followed by incubation with anti-digoxigenin-AP (1:10,000 in blocking solution) for 30 min and two washes with the Washing buffer, 15 min each. Membranes were finally rinsed in the Detection buffer [0.1 M Tris-HCl, pH 9.5 (Thermo Fisher Scientific, cat# BP152-1) and 0.1 M NaCl] for 5 min and chemiluminescence was detected using the CDP-Star reagent and an Odyssey imaging system (LI-COR Biosciences).

Immunoblotting

Cells grown in 6-well plates were washed three times with ice-cold 1×PBS and proteins were extracted using 100-200 μl/well of RIPA lysis buffer (Santa Cruz Biotechnology; cat# sc-364162) supplemented with 1 mM PMSF (New England Biolabs, cat# 8553S) and the recommended amount of cOmplete EDTA-free protease inhibitor cocktail (Roche, cat# 4693132001). Protein concentrations were determined using a Pierce BCA Protein Assay Kit.

Protein samples (10-25 µg) were then incubated at 95°C for 5 min in 1×Laemmli sample buffer (Bio-Rad; cat# 1610747), chilled on ice and separated by 4-20% gradient SDS-PAGE (Bio-Rad; cat# 4561096). The proteins were transferred from the gels to nitrocellulose membranes using a Trans-Blot Turbo Transfer System and analyzed using appropriate primary and secondary antibodies (see Supplementary Data 5). Protein bands were detected using an Odyssey imaging system and quantified using the LI-COR Image Studio software (LI-COR Biosciences).

Co-immunoprecipitation and RNA immunoprecipitation

2.0×10^6 A2lox ESCs were plated in 10 cm gelatinized dishes in 10 ml of 2i medium and immediately transfected with 500 pmol of an appropriate siRNA premixed with 27 µl of Lipofectamine 2000 and 1.5 ml of Opti-MEM I. 48 hours post transfection cells were washed three times with ice-cold 1×PBS and lysed in 600-700 µl of co-IP/RIP lysis buffer containing 10 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.5% NP40/IGEPAL CA-630 (Sigma, I8896) and the recommended amount of cOmplete EDTA-free protease inhibitor cocktail at 4°C for 30 min. In RIP experiments, co-IP/RIP lysis buffer was additionally supplemented with 100 units/ml of murine RNase inhibitor. The lysates were centrifuged at 16,000×g for 10 min at 4°C and We used 200-250 µl aliquots of the clarified lysate per individual co-IP/RIP experiment and stored 50 µl aliquots as input controls. The co-IP/RIP aliquots were mixed with 50 µl of Dynabeads protein G beads (Thermo Fisher Scientific, cat# 10003D) preloaded with 5 µg of protein-specific antibodies (Supplementary Data 5) or a non-immune rabbit IgG control (Thermo Fisher Scientific, cat# 10500C). Lysates were incubated with rotation at 4°C overnight. In some experiments, lysates were supplemented with 25 units/ml of benzonase (Merck, cat# 70664-3) before mixing them with the beads. Beads were washed 3 times with 200 µl PBS and 0.5% Tween 20 and bead-associated proteins and RNAs were eluted using 1×Laemmli sample buffer or TRIzol and analyzed by immunoblotting or RT-qPCR, respectively.

RNA-Seq

For RNA-Seq, A2lox cells were transfected with appropriate siRNAs as described above. Total RNAs were extracted 48 hours post transfection using TRIzol Plus RNA Purification Kit (Thermo Fisher Scientific cat# 12183555). RNAs were eluted in nuclease-free water, quality-controlled using a Bioanalyzer (Agilent) and hybridized with oligo(dT) magnetic beads to isolate the poly(A) RNA fraction used for subsequent library preparation steps. Stranded mRNA sequencing libraries were prepared using the TruSeq Stranded mRNA Library Preparation Kit (Illumina cat## RS-122-2101 and RS-122-2102). Purified libraries were qualified on an Agilent Technologies 2200 TapeStation using a D1000 ScreenTape assay (cat## 5067-5582 and 5067-5583). The molarity of adapter-modified molecules was defined by quantitative PCR using the Kapa Library Quant Kit (Kapa Biosystems; cat# KK4824). Individual libraries were normalized to 10 nM and equal volumes were pooled in preparation for Illumina sequence analysis. Sequencing libraries (25 pM) were chemically denatured and applied to an Illumina HiSeq v4 single read flow cell using an Illumina cBot. Hybridized molecules were clonally amplified and annealed to sequencing primers with reagents from a HiSeq SR Cluster Kit v4-cBot (Illumina; cat# GD-401-4001). Following transfer of the flowcell to a HiSeq 2500 instrument (Illumina; cat## HCSv2.2.38 and RTA v1.18.61), a 50-cycle single-read sequence run was performed using HiSeq SBS Kit v4 sequencing reagents (Illumina; cat# FC-401-4002). All library preparation and sequencing steps were carried out by the Huntsman Cancer Institute High-Throughput Genomics facility, University of Utah, USA.

3'RNA-Seq

To characterize global changes in cleavage/polyadenylation patterns, aliquots of total RNA samples prepared as described in the RNA-Seq section were additionally analyzed using 3'-proximal RNA-Seq (3'RNA-Seq). In this case, sequencing-ready libraries were produced using a QuantSeq 3' mRNA-Seq Library Prep Kit REV (Lexogen, cat# 016.24) following standard procedures, as outlined in the corresponding user guide (Lexogen; https://www.lexogen.com/wp-content/uploads/2018/08/015UG009V0241_QuantSeq_Illumina.pdf) using 200 ng of total RNA as input and using indexed primers for multiplexing. Finished libraries were quality-controlled using a Bioanalyzer (Agilent), using the High Sensitivity DNA assay. Library concentrations were determined using a Qubit dsDNA HS assay (Thermo Fisher scientific, cat# Q32851) and pooled for sequencing based on these quantifications. Sequencing was performed using an Illumina HiSeq2500 (v4) with SR75 High Output at the Vienna Biocenter Core Facilities. A custom sequencing primer (CSP) was used to sequence QuantSeq REV libraries. All library preparation and sequencing steps were carried out by the Lexogen GmbH service team, Austria.

RAP-Seq

RNA Antisense Purification (RAP) of formaldehyde-crosslinked samples was performed in principle as described [50]. 3.5×10^6 A2lox ESCs were plated in 10 cm gelatinized dishes in 10 ml of 2i medium and immediately transfected with 500 pmol of siRNAs premixed with 27 µl of Lipofectamine 2000 and 1.5 ml of Opti-MEM I. Medium was replaced once 24 hours post transfection and the culture was incubated for another 24 hours.

The cells (~8×10^6) were then washed once with 10 ml PBS and cross-linked with 7 ml of prewarmed 2% formaldehyde freshly diluted in PBS form 16% stock (Thermo Fischer Scientific, cat# 28908) for 10 min at 37°C with gentle rocking. Formaldehyde was quenched by adding 2.5 M glycine (Sigma, cat# G8898-500G) to a final concentration of 500 mM and incubating the plate at 37 °C for 5 min. Cells were then washed 3 times with cold PBS and scrapped off the plate in 2 ml of ice-cold Scraping Buffer [1× PBS and 0.5% DNase/RNase-free BSA (Thermo Fischer Scientific, cat# BP8805)], centrifuged at 1000×g at 4°C for 5 min, resuspended in hypotonic cell lysis buffer [10 mM HEPES pH 7.5 (Thermo Fischer Scientific, cat# 15630056), 20 mM KCl (Sigma, cat# P9541-1KG), 1.5 mM $MgCl_2$ (Sigma, cat# M8266-1KG), 0.5 mM EDTA (Thermo Fischer Scientific, cat# R1021), 1 mM tris(2-carboxyethyl)phosphine (TCEP) (Sigma, cat# 75259-1G), and 0.5 mM PMSF] and homogenized by douncing ~20 times with microtube pestles (STARLAB, cat# I1415-5390).

The lysates were centrifuged at 3,300×g for 7 min at 4°C and the pellets containing nuclei were resuspended in 1 ml of GuSCN Hybridization Buffer (20 mM Tris-HCl pH 7.5, 7 mM EDTA, 3 mM EGTA (Sigma, cat# E3889-10G), 150 mM LiCl (Sigma, cat# 62476-100G-F), 1% NP-40 (Sigma, cat# I8896-100ML), 0.2% N-lauroylsarcosine (Sigma, cat# L7414-10ML), 0.1% sodium deoxycholate (Sigma, cat# D6750-25G), 3M guanidine thiocyanate (Sigma, cat# G9277-100G), and 2.5 mM TCEP). We solubilized chromatin and fragmented RNA by sonicating the samples for 8 min using a Sonics Vibra-Cell VC130 Ultrasonic Processor equipped with a microtip, with pulser set to 10 seconds and the amplitude to 20. Lysates were centrifuged at 16,000×g for 10 min at 4°C and the supernatants were pre-cleared by incubating them for 30 min with MyONE Streptavidin C1 magnetic beads (100 µl original volume, compacted to 25 µl in GuSCN Hybridization Buffer; Thermo Fischer Scientific, cat# 65001) followed by magnetic separation in a DynaMag-2 rack (Thermo Fischer Scientific, cat# 12321D). Small aliquots (~10 µl) of pre-cleared lysates were saved and used later as RNA input controls.

For RAP, pre-cleared lysates from 5×10^6 cells were hybridized with 50 pmol of biotinylated DNA oligonucleotide probe against U1 snRNA (Supplementary Data 7) at 37°C for 2.5 hours with shaking at 1,200 rpm in a Thermomixer Compact (Eppendorf). The mixtures were then combined with MyONE Streptavidin C1 magnetic beads (500 µl original volume, compacted to 125 µl in GuSCN Hybridization Buffer) and incubated at 37 °C for 30 minutes with shaking. The beads were washed at 45°C with six changes of 500 µl GuSCN Wash Buffer (20 mM Tris-HCl pH 7.5, 10 mM EDTA, 1% NP-40, 0.2% N-lauroylsarcosine, 0.1% sodium deoxycyolate, 3 M guanidine thiocyanate, and 2.5 mM TCEP). We then washed the beads once in 500 µl of RNase H Elution Buffer (50 mM Tris-HCl pH 7.5, 75 mM NaCl, 3 mM $MgCl_2$, 0.125% N-lauroylsarcosine, 0.025% sodium deoxycholate, 2.5 mM TCEP) and once in 100 µl of RNase H Elution Buffer. The beads were subsequently resuspended in 55 µl RNase H Elution Buffer mixed with 7.5 µl RNase H (5 units/µl; New England Biolabs, cat# M0297S) and incubated at 37°C for 30 min with shaking to digest ssDNA-RNA hybrids and release U1-associated RNAs. The resultant eluates were stored on ice. Second elution step was performed by resuspending the beads in 62.5 µl GuSCN Hybridization Buffer and shaking for 5 min at 37°C. The first and second eluates were then combined.

To reverse crosslinks, the combined eluates and RNA inputs were mixed with 312.5 µl NLS Elution Buffer (20 mM Tris-HCl pH 7.5, 10 mM EDTA, 2% N-lauroylsarcosine, 2.5 mM TCEP), 50 µl 5 M NaCl, and 12.5 µl Proteinase K (Thermo Fischer Scientific, cat# EO0491) and incubated at 60°C for 2 hours. RNAs were then purified by mixing them with 40 µl of Dynabeads MyOne Silane beads (Thermo Fischer Scientific, cat# 37002D) pre-rinsed in RLT buffer (QIAGEN, cat# 79216) and resuspended in 50 µl 5 M NaCl. The suspensions were supplemented with 550 µl of 100% isopropanol, incubated for 2 min at room temperature, and magnetically separated. The beads were washed twice with 600 µl 70% ethanol and dried for 10 minutes. RNAs were eluted from the beads in 25 µl of nuclease-free water and treated with 2 units of TURBO DNAse in 1× TURBO DNAse buffer for 10 min at 37°C, without removing the beads from the tubes. The RNAs were then bound to the beads once again by adding 87.5 µl RLT and 112.5 µl isopropanol. The beads were washed twice in 70% ethanol, air-dried and RNAs were eluted from the beads in 25 µl of nuclease-free water.

RNAs were then processed using a NEBNext® rRNA Depletion Kit (New England Biolabs, cat# E6350S) as recommended. RNA-Seq libraries were generated using NEBNext® Ultra8482 II Directional RNA Library Preparation kit (New England Biolabs, cat# E7765S; following the protocol for rRNA Depleted FFPE/Strongly fragmented RNA). Individual libraries were normalized using Qubit, and their size profile was analyzed using TapeStation 4200. Individual libraries were normalized and pooled together accordingly. The pooled library was diluted to ~10 nM for storage. The 10 nM library was denatured and further diluted prior to loading on the sequencer. Paired-end sequencing was performed using a HiSeq4000 75bp platform (Illumina, HiSeq 3000/4000 PE Cluster Kit and 150 cycle SBS Kit). All library sequencing steps were carried out by the Oxford Genomics Centre, University of Oxford, UK.

Bioinformatics

All analyses were carried out using mm10 UCSC mouse genome and transcriptome files from Illumina (https://support.illumina.com/sequencing/sequencing_software/igenome.html) and UCSC Genome Browser (http://genome.ucsc.edu/). Canonical UCSC transcripts were used for most of the analyses (knownCanonical UCSC transcripts). Genomic intervals were analyzed using Bedtools or custom R-scripts. Duplicated features with identical genome positions and gene names were removed from the analyses.

For differential gene expression analyses, RNA-Seq reads were aligned with HISAT2 [74] using an mm10 UCSC-based genome index and a list of known splice junctions derived

from the UCSC-based mm10 genes.gtf file (ftp://igenome:G3nom3s4u@ussd-ftp.illumina.com/Mus_musculus/UCSC/mm10/Mus_musculus_UCSC_mm10.tar.gz). The alignment was done as follows:

hisat2 -p <n_threads> --rna-strandness F --known-splicesite-infile <hisat2_known_splice_sites.txt> -x <hisat2_genome_index> -U file1.fastq -S file1.sam

HISAT2-mapped reads were converted to BAM format using SAMtools [75] and assigned to annotated exons from the genes.gtf file using the featureCounts function of the Rsubread R/Bioconductor package [76] in a strand-specific manner. Differentially expressed genes were then identified using the edgeR package with the estimateGLMRobustDisp function [77,78]. GO-term enrichment was calculated using the goseq package [79] with gene lengths taken into account. Venn diagrams and gene expression heat maps were generated using VennDiagram (https://cran.r-project.org/web/packages/VennDiagram/) and pheatmap packages (https://cran.r-project.org/web/packages/pheatmap/), respectively. RNA-Seq coverage metaplots were prepared using ngs.plot [80].

Relative intron coverage (*RIC*) statistic was calculated as:

$$RIC = I/E \quad (1)$$

where I is the total number of intronic reads and reads spanning junctions between the intron and the adjacent exons by ≥10 nt and E is the number of reads matching the adjacent exons and their splice junction. Reads were assigned to the I and E intervals using Bedtools [81]. Statistical significance of *RIC* changes was assessed by two-tailed Fisher's exact test comparison of I and E values between two experimental conditions. Entries with $I<5$ and $E<10$ in both conditions were excluded from the analysis. False discovery rate (FDR) was calculated by adjusting the resultant *p*-values using the Benjamini-Hochberg method.

To analyze changes in cleavage/polyadenylation patterns, 3'-proximal RNA-Seq data were aligned to mm10 genome using Bowtie2 [82] with trimming the first 12 nt to remove poly(A) tail-derived sequences:

bowtie2 --fast --trim5 12 -N 1 -p <n_threads> -x <Bowtie2_genome_index> -U file1.fastq -S file1.sam

Reads with high probability of being primed internally rather than at bona fide poly(A) tails were identified by inspecting corresponding genomic sequences. If 10 consecutive adenosines (with one mismatch allowed) were found within a 20-nt genomic window preceding the read, the read was discarded. The first 5'-terminal nucleotide of the remaining reads mapping to the genome was considered to match a CSs. Individual CSs were then clustered by merging positions spaced by ≤10 nt across all experimental samples. Clusters containing ≥3 reads in at least one sample were kept for further analyses. Clusters were allocated to known intronic and exonic features from the mm10 UCSC annotation using Bedtools.

Incidence of PAS hexamers in a 50 nt window bounded by 40 nt upstream and 10 nt downstream of the middle of CS clusters was calculated using a custom Python script. Cleavage/polyadenylation clusters were considered novel if their middle was >50 nt away from annotated cleavage/polyadenylation sites from the polyA_DB3 database [48] converted from mm9 to mm10 coordinates using USCS Genome Browser liftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver).

Relative cleavage/polyadenylation site efficiency (*RCE*) was calculated as:

$$RCE = \frac{N_k}{\sum_{i=0}^{n} N_i} \quad (2)$$

where N_k is the number of reads matching the cleavage/polyadenylation cluster k and n is the total number of reads mapping to cleavage/polyadenylation clusters in the same gene. Statistical significance of changes in cleavage/polyadenylation cluster usage was assessed using two-tailed Fisher's exact test by comparing N_k and $(\sum_{i=0}^{n} N_i) - N_k$ values between

experimental conditions. FDR was calculated using the Benjamini-Hochberg method. We used RCE fold change and FDR values to shortlist significantly regulated CSs. In many cases, we aggregated *RCE* values for specific genomic ranges (e.g. first introns or 3'UTRs; Figs. 2c and 4b, c and Supplementary Figs. 4d and 10c) and plotted a normalized difference in this statistic between experimental (e) and control (c) samples:

$$\Delta RCE_{norm} = \frac{RCE_e - RCE_c}{RCE_e + RCE_c} \quad (3)$$

To generate metaplots for 3'RNA-Seq data (Supplementary Fig. 8a, b), genomic regions of interest were split into equally sized bins and a normalized change in 3'-proximal read coverage was calculated for each bin as follows:

$$3'RC_{norm} = \frac{RPM_e - RPM_c}{RPM_e + RPM_c} \quad (4)$$

where RPM_e and RPM_c are bin-specific coverage data for experimental and control conditions. The bin-specific $3'RC_{norm}$ values were then averaged across different genes and plotted after smoothing with Loess function in R (span = 0.15). A similar approach was used to prepare Supplementary Fig. 8c where we compared untransformed $3'RC_{norm}$ values for 3'UTRs of individual genes. In cases where metaplots for sense and antisense strands had to be shown on the same graph, the antisense strand data were multiplied by -1.

For RAP-Seq data analysis reads were aligned with Bowtie2 using an mm10 UCSC-based bowtie2 genome index as follows:

bowtie2 --fast -N 1 -p <n_threads> -x <Bowtie2_genome_index> -1 file1_1.fastq -2 file1_2.fastq -S file1.sam

Aligned fragments were sorted and converted to genomic intervals using `pairedBamToBed12` tool (https://github.com/Population-Transcriptomics/pairedBamToBed12). Fragments with mapping quality <30 were discarded. Piranha peak caller [51] was used to identify RAP-Seq clusters interacting with U1 snRNA using corresponding input samples as a background:

Piranha -o <output_file> -p 0.01 -a 0.85 -s -l -b 100 -i 100 RAP_1.bed Input_1.bed

Only RAP-Seq clusters present in both replicates were considered for further analysis. Cluster density in specific genomic intervals was calculated using Bedtools. Alternatively, RAP-Seq signal was normalized to input using bamCompare function of the Deeptools package [83] as follows:

bamCompare -b1 RAP1_merged.bam -b2 Input1_merged.bam --normalizeUsing RPKM --scaleFactorsMethod None --numberOfProcessors <n_threads> --binSize 25 --operation log2 --smoothLength 75 -o log2ratio25_RAP1.bw

and visualized using IGV [84].

To prepare metaplots for RAP-Seq data, genomic regions of interest were divided into 100 bins and the bamCompare-processed values were averaged for each bin using Bedtools and plotted as mean ±SEM.

PhastCons data for placental mammals [52] were downloaded from UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenpath/mm10/phastCons60way/mm10.60way.phastCons60wayPlacental.bw) and average PhastCons scores were calculated for 50-nt windows bounded by 40 nt upstream and 10 nt downstream of the middle of CS clusters.

RepeatMasker data for retrotransposable elements (RTEs) were retrieved from UCSC Genome Browser. RTE consensus sequences were obtained from https://www.girinst.org/repbase/. To generate RTE density metaplots, 2 kb windows centered on the middle of CS clusters were divided into 100 bins and SINE, LINE and LTR coverage for each bin was calculated using Bedtools and plotted as mean ±SEM. Divergence of individual RTEs from consensus sequence was assessed using RepeatMasker milliDiv statistic (base mismatches in parts per thousand; http://www.repeatmasker.org). Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) and EMBOSS Matcher

(https://www.ebi.ac.uk/Tools/psa/emboss_matcher/nucleotide.html) were used to generate multiple and pairwise DNA sequence alignments, respectively. Strength of putative U1-binding motifs was estimated using MaxEntScan::score5ss [85].

Statistical analyses

Unless stated otherwise, all statistical procedures were performed in R, and experimental data were averaged from at least three experiments and shown with error bars representing SD. Data obtained from RT-qPCR and immunoblot quantifications, were typically analyzed using a two-tailed Student's t-test assuming unequal variances. Correlation analyses were done using Pearson's product-moment and Spearman and Kendall's rank correlation methods, as specified in the text. Genome-wide data were typically compared using two-tailed Wilcoxon rank sum test (for non-paired count data), two-tailed Wilcoxon signed rank test (for paired count data) or two-tailed Fisher's exact test (for categorical data). Where necessary, p-values were adjusted for multiple testing using Benjamini-Hochberg correction (FDR). Numbers of experimental replicates, *p*-values and the tests used are indicated in the Figures and/or Figure legends.

Acknowledgements

We thank Carolina Barcellos Machado, Georgii Bazykin, Fursham Hamid, Michael Kyba, Ivo Lieberam, Stefan Mockenhaupt, Karen Yap, Feng Zhang, and Anna Zhuravskaya for reagents and helpful discussions. We are also grateful to Snezhka Oliferenko for valuable comments on the manuscript. This work was supported by the Biotechnology and Biological Sciences Research Council (BB/M001199/1, BB/M007103/1 and BB/R001049/1) and European Commission (H2020-MSCA-RISE-2016; Project ID 734791).

Author Contributions

Y.A.K. designed and conducted the experiments, analyzed the data and wrote the paper. E.V.M. designed experiments, analyzed the data and wrote the paper.

Declaration of Interests

The authors declare no competing interests.

Data availability

The RNA-Seq, 3'RNA-Seq and RAP-Seq data generated in this study are available from ArrayExpress (E-MTAB-7626, E-MTAB-7635). Publicly available sequencing data used in our study are summarized in Supplementary Data 5. The source data underlying Figs. 1b, c, e-g, 2d, 3b-e, 4d-f, 5e and 6a, c and Supplementary Figs. 1a, b, 2c, 5c, 6a-e, 7e, f, 8d, e, 10d-f, 11a and 12a-d are provided as a Source Data file. All other data are available from the authors.

Code availability

Computer code used in this study is described in the Methods and Supplementary Data 5.

References

1 Maniatis, T. & Reed, R. An extensive network of coupling among gene expression machines. *Nature* **416**, 499-506, doi:10.1038/416499a (2002).

2 Moore, M. J. & Proudfoot, N. J. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* **136**, 688-700, doi:10.1016/j.cell.2009.02.001 (2009).

3 Skalska, L., Beltran-Nebot, M., Ule, J. & Jenner, R. G. Regulatory feedback from nascent RNA to chromatin and transcription. *Nat Rev Mol Cell Biol* **18**, 331-337, doi:10.1038/nrm.2017.12 (2017).

4 Saldi, T., Cortazar, M. A., Sheridan, R. M. & Bentley, D. L. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *J Mol Biol* **428**, 2623-2635, doi:10.1016/j.jmb.2016.04.017 (2016).

5 Bresson, S. & Tollervey, D. Surveillance-ready transcription: nuclear RNA decay as a default fate. *Open Biol* **8**, doi:10.1098/rsob.170270 (2018).

6 Hsin, J. P. & Manley, J. L. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev* **26**, 2119-2137, doi:10.1101/gad.200303.112 (2012).

7 Jensen, T. H., Jacquier, A. & Libri, D. Dealing with pervasive transcription. *Mol Cell* **52**, 473-484, doi:10.1016/j.molcel.2013.10.032 (2013).

8 Gonatopoulos-Pournatzis, T. & Cowling, V. H. Cap-binding complex (CBC). *Biochem J* **457**, 231-242, doi:10.1042/BJ20131214 (2014).

9 Muller-McNicoll, M. & Neugebauer, K. M. Good cap/bad cap: how the cap-binding complex determines RNA fate. *Nat Struct Mol Biol* **21**, 9-12, doi:10.1038/nsmb.2751 (2014).

10 Gruber, J. J. *et al.* Ars2 links the nuclear cap-binding complex to RNA interference and cell proliferation. *Cell* **138**, 328-339, doi:10.1016/j.cell.2009.04.046 (2009).

11 Hallais, M. *et al.* CBC-ARS2 stimulates 3'-end maturation of multiple RNA families and favors cap-proximal processing. *Nat Struct Mol Biol* **20**, 1358-1366, doi:10.1038/nsmb.2720 (2013).
12 Andersen, P. R. *et al.* The human cap-binding complex is functionally connected to the nuclear RNA exosome. *Nat Struct Mol Biol* **20**, 1367-1376, doi:10.1038/nsmb.2703 (2013).
13 Schulze, W. M., Stein, F., Rettel, M., Nanao, M. & Cusack, S. Structural analysis of human ARS2 as a platform for co-transcriptional RNA sorting. *Nat Commun* **9**, 1701, doi:10.1038/s41467-018-04142-7 (2018).
14 Gruber, J. J. *et al.* Ars2 promotes proper replication-dependent histone mRNA 3' end formation. *Mol Cell* **45**, 87-98, doi:10.1016/j.molcel.2011.12.020 (2012).
15 Grigg, S. P., Canales, C., Hay, A. & Tsiantis, M. SERRATE coordinates shoot meristem function and leaf axial patterning in Arabidopsis. *Nature* **437**, 1022-1026, doi:10.1038/nature04052 (2005).
16 Sabin, L. R. *et al.* Ars2 regulates both miRNA- and siRNA- dependent silencing and suppresses RNA virus infection in Drosophila. *Cell* **138**, 340-351, doi:10.1016/j.cell.2009.04.045 (2009).
17 Gornemann, J., Kotovic, K. M., Hujer, K. & Neugebauer, K. M. Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. *Mol Cell* **19**, 53-63, doi:10.1016/j.molcel.2005.05.007 (2005).
18 Lewis, J. D., Izaurralde, E., Jarmolowski, A., McGuigan, C. & Mattaj, I. W. A nuclear cap-binding complex facilitates association of U1 snRNP with the cap-proximal 5' splice site. *Genes Dev* **10**, 1683-1698 (1996).
19 Pabis, M. *et al.* The nuclear cap-binding complex interacts with the U4/U6.U5 tri-snRNP and promotes spliceosome assembly in mammalian cells. *RNA* **19**, 1054-1063, doi:10.1261/rna.037069.112 (2013).
20 Elahi, S. *et al.* The RNA binding protein Ars2 supports hematopoiesis at multiple levels. *Exp Hematol* **64**, 45-58 e49, doi:10.1016/j.exphem.2018.05.001 (2018).
21 Olejniczak, S. H., La Rocca, G., Gruber, J. J. & Thompson, C. B. Long-lived microRNA-Argonaute complexes in quiescent cells can be activated to regulate mitogenic responses. *Proc Natl Acad Sci U S A* **110**, 157-162, doi:10.1073/pnas.1219958110 (2013).
22 O'Sullivan, C. S. *et al.* ARS2 is required for retinal progenitor cell S-phase progression and Muller glial cell fate specification. *Biochem Cell Biol*, doi:10.1139/bcb-2018-0250 (2019).
23 O'Sullivan, C. *et al.* Mutagenesis of ARS2 Domains To Assess Possible Roles in Cell Cycle Progression and MicroRNA and Replication-Dependent Histone mRNA Biogenesis. *Mol Cell Biol* **35**, 3753-3767, doi:10.1128/MCB.00272-15 (2015).
24 Andreu-Agullo, C., Maurin, T., Thompson, C. B. & Lai, E. C. Ars2 maintains neural stem-cell identity through direct transcriptional activation of Sox2. *Nature* **481**, 195-198, doi:10.1038/nature10712 (2011).
25 Wilson, M. D. *et al.* ARS2 is a conserved eukaryotic gene essential for early mammalian development. *Mol Cell Biol* **28**, 1503-1514, doi:10.1128/MCB.01565-07 (2008).
26 Golling, G. *et al.* Insertional mutagenesis in zebrafish rapidly identifies genes essential for early vertebrate development. *Nat Genet* **31**, 135-140, doi:10.1038/ng896 (2002).
27 Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18**, 18-30, doi:10.1038/nrm.2016.116 (2017).

28 Shi, Y. & Manley, J. L. The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site. *Genes Dev* **29**, 889-897, doi:10.1101/gad.261974.115 (2015).

29 Neve, J., Patel, R., Wang, Z., Louey, A. & Furger, A. M. Cleavage and polyadenylation: Ending the message expands gene regulation. *RNA Biol* **14**, 865-890, doi:10.1080/15476286.2017.1306171 (2017).

30 Proudfoot, N. J. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science* **352**, aad9926, doi:10.1126/science.aad9926 (2016).

31 Kaida, D. *et al.* U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664-668, doi:10.1038/nature09479 (2010).

32 Berg, M. G. *et al.* U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150**, 53-64, doi:10.1016/j.cell.2012.05.029 (2012).

33 Oh, J. M. *et al.* U1 snRNP telescripting regulates a size-function-stratified human genome. *Nat Struct Mol Biol* **24**, 993-999, doi:10.1038/nsmb.3473 (2017).

34 Ntini, E. *et al.* Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nat Struct Mol Biol* **20**, 923-928, doi:10.1038/nsmb.2640 (2013).

35 Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360-363, doi:10.1038/nature12349 (2013).

36 Chiu, A. C. *et al.* Transcriptional Pause Sites Delineate Stable Nucleosome-Associated Premature Polyadenylation Suppressed by U1 snRNP. *Mol Cell* **69**, 648-663 e647, doi:10.1016/j.molcel.2018.01.006 (2018).

37 Martello, G. & Smith, A. The nature of embryonic stem cells. *Annu Rev Cell Dev Biol* **30**, 647-675, doi:10.1146/annurev-cellbio-100913-013116 (2014).

38 Young, R. A. Control of the embryonic stem cell state. *Cell* **144**, 940-954, doi:10.1016/j.cell.2011.01.032 (2011).

39 Dunn, S. J., Li, M. A., Carbognin, E., Smith, A. & Martello, G. A common molecular logic determines embryonic stem cell self-renewal and reprogramming. *EMBO J* **38**, doi:10.15252/embj.2018100003 (2019).

40 Hubbard, K. S., Gut, I. M., Lyman, M. E. & McNutt, P. M. Longitudinal RNA sequencing of the deep transcriptome during neurogenesis of cortical glutamatergic neurons from murine ESCs. *F1000Res* **2**, 35, doi:10.12688/f1000research.2-35.v1 (2013).

41 Guo, G. *et al.* Serum-Based Culture Conditions Provoke Gene Expression Variability in Mouse Embryonic Stem Cells as Revealed by Single-Cell Analysis. *Cell Rep* **14**, 956-965, doi:10.1016/j.celrep.2015.12.089 (2016).

42 Ying, Q. L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519-523, doi:10.1038/nature06968 (2008).

43 Kalkan, T. *et al.* Tracking the embryonic stem cell transition from ground state pluripotency. *Development* **144**, 1221-1234, doi:10.1242/dev.142711 (2017).

44 Ogawa, K. *et al.* Activin-Nodal signaling is involved in propagation of mouse embryonic stem cells. *J Cell Sci* **120**, 55-65, doi:10.1242/jcs.03296 (2007).

45 Moyses-Oliveira, M. *et al.* Inactivation of AMMECR1 is associated with growth, bone, and heart alterations. *Hum Mutat* **39**, 281-291, doi:10.1002/humu.23373 (2018).

46 Burroughs, A. M. & Aravind, L. A highly conserved family of domains related to the DNA-glycosylase fold helps predict multiple novel pathways for RNA modifications. *RNA Biol* **11**, 360-372, doi:10.4161/rna.28302 (2014).

47 Tsai, T. C., Lee, Y. L., Hsiao, W. C., Tsao, Y. P. & Chen, S. L. NRIP, a novel nuclear receptor interaction protein, enhances the transcriptional activity of nuclear receptors. *J Biol Chem* **280**, 20000-20009, doi:10.1074/jbc.M412169200 (2005).

48 Wang, R., Nambiar, R., Zheng, D. & Tian, B. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res* **46**, D315-D319, doi:10.1093/nar/gkx1000 (2018).

49 Zheng, G. X., Do, B. T., Webster, D. E., Khavari, P. A. & Chang, H. Y. Dicer-microRNA-Myc circuit promotes transcription of hundreds of long noncoding RNAs. *Nat Struct Mol Biol* **21**, 585-590, doi:10.1038/nsmb.2842 (2014).

50 Engreitz, J. M. *et al.* RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. *Cell* **159**, 188-199, doi:10.1016/j.cell.2014.08.018 (2014).

51 Uren, P. J. *et al.* Site identification in high-throughput RNA-protein interaction data. *Bioinformatics* **28**, 3013-3020, doi:10.1093/bioinformatics/bts569 (2012).

52 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).

53 Hancks, D. C. & Kazazian, H. H., Jr. Roles for retrotransposon insertions in human disease. *Mob DNA* **7**, 9, doi:10.1186/s13100-016-0065-9 (2016).

54 Elbarbary, R. A., Lucas, B. A. & Maquat, L. E. Retrotransposons as regulators of gene expression. *Science* **351**, aac7247, doi:10.1126/science.aac7247 (2016).

55 Kramerov, D. A. & Vassetzky, N. S. SINEs. *Wiley Interdiscip Rev RNA* **2**, 772-786, doi:10.1002/wrna.91 (2011).

56 Zavolan, M. & Kanitz, A. RNA splicing and its connection with other regulatory layers in somatic cell reprogramming. *Curr Opin Cell Biol* **52**, 8-13, doi:10.1016/j.ceb.2017.12.002 (2018).

57 Han, H. *et al.* MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**, 241-245, doi:10.1038/nature12270 (2013).

58 Corsini, N. S. *et al.* Coordinated Control of mRNA and rRNA Processing Controls Embryonic Stem Cell Pluripotency and Differentiation. *Cell Stem Cell* **22**, 543-558 e512, doi:10.1016/j.stem.2018.03.002 (2018).

59 Lu, X. *et al.* SON connects the splicing-regulatory network with pluripotency in human embryonic stem cells. *Nat Cell Biol* **15**, 1141-1152, doi:10.1038/ncb2839 (2013).

60 Nudler, E. & Gottesman, M. E. Transcription termination and anti-termination in E. coli. *Genes Cells* **7**, 755-768 (2002).

61 Laubinger, S. *et al.* Dual roles of the nuclear cap-binding complex and SERRATE in pre-mRNA splicing and microRNA processing in Arabidopsis thaliana. *Proc Natl Acad Sci U S A* **105**, 8795-8800, doi:10.1073/pnas.0802493105 (2008).

62 Raczynska, K. D. *et al.* The SERRATE protein is involved in alternative splicing in Arabidopsis thaliana. *Nucleic Acids Res* **42**, 1224-1244, doi:10.1093/nar/gkt894 (2014).

63 Iasillo, C. *et al.* ARS2 is a general suppressor of pervasive transcription. *Nucleic Acids Res* **45**, 10229-10241, doi:10.1093/nar/gkx647 (2017).

64 Robbez-Masson, L. & Rowe, H. M. Retrotransposons shape species-specific embryonic stem cell gene expression. *Retrovirology* **12**, 45, doi:10.1186/s12977-015-0173-5 (2015).

65 Cost, G. J., Golding, A., Schlissel, M. S. & Boeke, J. D. Target DNA chromatinization modulates nicking by L1 endonuclease. *Nucleic Acids Res* **29**, 573-577 (2001).

66 Klawitter, S. *et al.* Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. *Nat Commun* **7**, 10286, doi:10.1038/ncomms10286 (2016).
67 Lee, J. Y., Ji, Z. & Tian, B. Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3'-end of genes. *Nucleic Acids Res* **36**, 5581-5590, doi:10.1093/nar/gkn540 (2008).
68 Attig, J. *et al.* Heteromeric RNP Assembly at LINEs Controls Lineage-Specific RNA Processing. *Cell* **174**, 1067-1081 e1017, doi:10.1016/j.cell.2018.07.001 (2018).
69 Goodier, J. L. Restricting retrotransposons: a review. *Mob DNA* **7**, 16, doi:10.1186/s13100-016-0070-z (2016).
70 Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* **11**, 345-355, doi:10.1038/nrg2776 (2010).
71 Iacovino, M. *et al.* Inducible cassette exchange: a rapid and efficient system enabling conditional gene expression in embryonic stem and primary cells. *Stem Cells* **29**, 1580-1588 (2011).
72 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823, doi:10.1126/science.1231143 (2013).
73 Scotto-Lavino, E., Du, G. & Frohman, M. A. 3' end cDNA amplification using classic RACE. *Nat Protoc* **1**, 2742-2745, doi:10.1038/nprot.2006.481 (2006).
74 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).
75 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
76 Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* **41**, e108, doi:10.1093/nar/gkt214 (2013).
77 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).
78 Zhou, X., Lindsay, H. & Robinson, M. D. Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res* **42**, e91, doi:10.1093/nar/gku310 (2014).
79 Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11**, R14, doi:10.1186/gb-2010-11-2-r14 (2010).
80 Shen, L., Shao, N., Liu, X. & Nestler, E. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* **15**, 284, doi:10.1186/1471-2164-15-284 (2014).
81 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).
82 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
83 Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165, doi:10.1093/nar/gkw257 (2016).
84 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).

85 Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377-394, doi:10.1089/1066527041410418 (2004).
86 Yap, K., Xiao, Y., Friedman, B. A., Je, H. S. & Makeyev, E. V. Polarizing the Neuron through Sustained Co-expression of Alternatively Spliced Isoforms. *Cell Rep* **15**, 1316-1328, doi:10.1016/j.celrep.2016.04.012 (2016).
87 Kalkan, T. *et al.* Complementary Activity of ETV5, RBPJ, and TCF3 Drives Formative Transition from Naive Pluripotency. *Cell Stem Cell* **24**, 785-801 e787, doi:10.1016/j.stem.2019.03.017 (2019).

Figure Legends

Figure 1. Srrt is required for mouse ESC maintenance

(**a**) Bioinformatics workflow used to identify putative regulators of mouse ESC identity.

(**b**) *Top*, immunoblot analysis of Srrt expression in mouse ESCs, cortical NSCs and cortical neurons prepared and cultured in vitro as described [86]. *Bottom*, Srrt protein expression was quantified from 3 independent experiments (mean±SD) and compared using a two-tailed t-test.

(**c**) *Top*, ESCs were transfected with a Srrt-specific siRNA mixture (siSrrt) or a non-targeting control siRNA (siCtrl) and Srrt knockdown efficiency was analyzed by immunoblotting 48 hours later. *Bottom*, the experiment was repeated twice (mean±SD) and the samples were compared using a two-tailed t-test. (b, c) Erk1/2 is a lane loading control.

(**d**) ESCs were transfected with siSrrt as in (c) and assayed for alkaline phosphatase (AP) activity. Note pronounced changes in morphology of colonies and individual cells and a decrease in the AP staining intensity. Scale bar, 100 μm.

(**e, f**) Colony assay data showing that (e) siSrrt does not change the overall number of ESC colonies but (f) significantly increases the fraction of flattened differentiated colonies compared to siCtrl. The assay was repeated 3 times (mean±SD) and analyzed by a two-tailed t-test.

(**g**) *Left*, RT-qPCR data showing that, while Srrt knockdown does not change expression of pluripotency markers Pou5f1, Sox2, Nanog and Zfp42/Rex1, it leads to significant downregulation of Nr0b1, Pecam1 and Zic2 and upregulation of early differentiation markers Etv4, Otx2 and Runx1 [39,43,87]. *Right*, targets strongly downregulated by siSrrt include additional examples of known ESC markers and factors with possible regulatory roles in proliferating cells [43-47]. All RT-qPCR experiments were done at least in triplicate and shown

as mean ±SD. The expression levels in siCtrl-treated samples were set to 1, and the *p*-values were calculated using a two-tailed t-test. Source data are provided as a Source Data file.

Figure 2. Srrt blocks cleavage/polyadenylation in first introns of many genes

(**a**) Srrt knockdown in mouse ESCs promotes utilization of cryptic CSs in first introns.

(**b**) Upregulated CSs tend to localize close to the 5' end of fist introns. (a, b) CSs with FC≥2 and FDR<0.05 were considered significantly regulated.

(**c**) Scatter plot showing that siSrrt-mediated activation of intronic CSs strongly correlates with downregulation of gene expression. Genes with significant changes in relative CS efficiency in first introns (FDR<0.05) and expression levels (FC≥1.5 and FDR<0.05) are shown in red. Other genes, gray.

(**d**) Examples of genes regulated by Srrt via intronic cleavage/polyadenylation. Read-per-million (rpm)-normalized RNA-Seq coverage plots are shown in gray, and rpm-normalized 3'RNA-Seq data are in red. Note simultaneous activation of CSs in first introns and a decrease in RNA-Seq and 3'RNA-Seq signals in the corresponding 3' untranslated regions (3'UTRs). Red arrowheads, CSs preceded by canonical polyadenylation signals (PASs), AATAAA or ATTAAA.

(**e**) RT-qPCR verification of the siSrrt effect on genes in (d) using primer pairs designed against sequences upstream or downstream of regulated iCSs. Gene-specific signals were normalized to Cnot4 housekeeping gene and the expression levels in siCtrl-treated sample were set to 1. Data were averaged from 3 experiments ±SD and compared by a two-tailed t-test. Source data are provided as a Source Data file.

Figure 3. Intronic cleavage/polyadenylation is required for *Ammecr1* regulation by Srrt

(**a**) *Top*, Ammecr1 wild-type (*WT*) intronic sequence regulated in response to Srrt knockdown. Canonical PAS motifs are highlighted in pink. Also shown are positions of CRISPR gRNAs used to generate the *ΔPAS* allele. Sequence deleted in *ΔPAS* is in lowercase. *Bottom*, Sanger sequence analysis of the ΔPAS Ammecr1 allele.

(**b**) PCR genotyping result comparing WT and ΔPAS ESCs.

(**c**) Passage-matched WT and ΔPAS ESC clones were treated with either siSrrt or siCtrl and the efficiency of Srrt knockdown was analyzed by RT-qPCR 48 hours later. Note that Srrt levels decrease to a comparable extent in both genetic backgrounds.

(**d, e**) The effect of siSrrt on the expression of Ammecr1 sequences (d) upstream and (e) downstream of the iCS in the *WT* (and the deleted intronic region in the *ΔPAS* allele). Note that deletion of the CS region in ΔPAS cells abolishes (d) siSrrt-induced upregulation of the truncated 5'-proximal transcript and (e) downregulation of the full-length isoform. Data in (c-e) were averaged from 3 experiments ±SD, normalized to the WT/siCtrl samples, and compared by a two-tailed t-test. Source data are provided as a Source Data file.

Figure 4. Srrt-mediated repression of iCSs depends on the CBC

(**a**) Workflow used to compare transcriptome-wide effects of siSrrt and an siRNA targeting Ncbp1.

(**b**) Scatter plot showing a correlation (Pearson's r=0.74, p=0) between the effects of siSrrt and siNcbp1 on CSs in first introns. Note that most iCSs significantly regulated by both siSrrt and siNcbp1 (FDR<0.05; red) show an increase in relative efficiency (top right quadrant).

(**c**) Scatter plot showing that, similar to siSrrt (Fig. 2c), siNcbp1-mediated activation of iCSs often coincides with downregulation of corresponding genes. Red, genes with significant changes in relative CS efficiency in first introns (FDR<0.05) and expression levels (FC≥1.5 and FDR<0.05) are shown in red. Gray, the rest of the genes.

(**d**) ESCs containing a human SRRT transgene (SRRT-Tg; *TRE-SRRT-r3'UTR*) or a control expression cassette (Control-Tg; *TRE-EGFP-r3'UTR*) were pretreated with 2 μg/ml Dox for 24 hours and transfected with siCtrl, siNcbp1 or siSrrt. Expression levels of the Ncbp1 and Srrt proteins were analyzed by immunoblotting 48 hours later. Note that, compared to siCtrl, siNcbp1 and siSrrt reduce the abundance of the corresponding proteins in both transgenic backgrounds. However, the combined Srrt/SRRT expression in the SRRT-Tg/siSrrt sample still exceeds the Srrt levels in Control-Tg/siCtrl. Erk1/2, lane loading control. *TRE*, Dox-inducible promoter; *r3'UTR*, recombinant 3'UTR from SV40 virus. For quantification of this experiment see Supplementary Fig. 10d.

(**e, f**) RT-qPCR analysis showing that (e) both siSrrt and siNcbp1 decrease transcriptional readthrough of iCS in the *Ammecr1* gene in the Control-Tg background. (f) Recombinant SRRT rescues the effect of siSrrt but not siNcbp1 in the SRRT-Tg cells suggesting that Ncbp1 is essential for Srrt-mediated repression of iCSs. Data in (e-f) were averaged from 3 experiments ±SD and compared by a two-tailed t-test. Source data are provided as a Source Data file.

Figure 5. Srrt stimulates U1 binding upstream of CSs in first introns

(**a**) Outline of the U1 RAP-Seq experiment.

(**b**) Boxplot of U1 RAP-Seq cluster coverage showing stronger binding of U1 snRNP in a 250-nt window upstream of Srrt-regulated iCSs than in a similarly sized window downstream of these sites in siCtrl-treated samples. Note that U1 binding efficiency is diminished following Srrt knockdown. *P*-values were calculated using a two-tailed Wilcoxon signed rank test. The box bounds represent the first and the third quartiles and the thick black lines at the bottom of the boxes show the medians. Since the distributions are skewed towards 0, only the

top whisker is evident, extending to 1.5× of the range between the third and the first quartiles (interquartile range). Open circles, outliers.

(**c**) Consistent with the data in (b), the 250-nt window upstream of Srrt-repressed CSs tends to contain stronger putative U1 binding motifs (measured as the maximum 5'ss MaxEnt value) than the 250-nt downstream window or similarly sized windows abutting CSs in the corresponding 3'UTRs. *P*-values were calculated using a two-tailed Wilcoxon rank sum test. Violin plot outlines show kernel density estimates of probability densities; open circles, the medians; and bounds of the black boxes, the first and the third quartiles. Whiskers extend from the first and the third quartile to the lowest and highest data points or, if there are outliers, 1.5× of the interquartile range. (b, c) iCSs were considered regulated if they were upregulated ≥2-fold, FDR<0.05 and their host gene was downregulated ≥1.5-fold, FDR<0.05 in response to siSrrt.

(**d**) Input-normalized RAP-Seq coverage profile and Piranha clusters (U1-1 and U1-2) showing strong interaction of U1 snRNP with at least two intronic positions preceding the Srrt-repressed CS in the *Ammecr1* gene in the siCtrl- but not siSrrt-treated ESCs. Sequences enriched in RAP-Seq v input are shown in black and those depleted are in gray. Primers used in the RT-qPCR validation experiment in (e) are shown at the bottom.

(**e**) RT-qPCR validation of RAP-Seq results using primer pairs matching U1 Piranha clusters in (b) and Supplementary Fig. 11e. Note that input-normalized signals are significantly higher in siCtrl U1 RAP samples than in their siSrrt-treated counterparts for the two regulated Ammecr1 clusters but not for a control cluster in the Ncbp2 pre-mRNA. Data were averaged from 3 experiments ±SD and compared by a two-tailed t-test. Source data are provided as a Source Data file.

Figure 6. Srrt-mediated readthrough of Ammecr1 iCS depends on telescripting

(**a**) Mouse ESCs were nucleofected with a control morpholino oligonucleotide (amoCtrl) or antisense morpholinos targeting either U1 or U2 snRNA (amoU1 and amoU2) for 8 hours and the effect of these treatments on the iCS readthrough was analyzed using RT-qPCR. Note that amoU1 leads to a robust decrease in the CS readthrough efficiency compared to amoCtrl and amoU2.

(**b**) Ammecr1-based minigene constructs used in telescripting assays in (C). *PSV40*, SV40 enhancer and early promoter; *E1*, the first exon of *Ammecr1* gene; *r3'UTR*, recombinant 3'UTR from SV40 virus.

(**c**) ESCs transiently transfected with wild-type (*WT*) or mutant (*Mut-5'ss*, *Mut-4motifs* or *ΔPAS*) minigenes from (b) were treated with either siCtrl or siSrrt and the Ammecr1 intronic CS efficiency was assayed as a ratio between downstream [mini_F2/mini_R2 primers in (b)] and upstream RT-qPCR signals [mini_F1/mini_R1 primers in (b)]. Note that Srrt stimulates CS readthrough in the WT minigene, similar to its effect on the endogenously encoded *Ammecr1*. Mutation of a single U1-binding motif corresponding to the 5'ss at the beginning of the first intron (*Mut-5'ss*) does not alter the minigene response to Srrt knockdown; however, mutation of 5'ss and three additional sites potentially interacting with U1 (*Mut-4motifs*) results in a constitutive cleavage/polyadenylation phenotype. Conversely, deletion of the two PAS motifs (*ΔPAS*) leads to constitutive readthrough. Data on (a) and (c) were averaged from 3 experiments ±SD and compared by a two-tailed t-test. Readthrough efficiencies in the amoCtrl and *WT*/siCtrl samples, respectively, were set to 1. Source data are provided as a Source Data file.

Figure 7. Regulated iCSs often appear as a result of retrotransposition

(**a**) Fisher's exact test showing that Srrt-regulated iCSs are less frequently conserved across placental mammals as compared to their 3'UTR counterparts.

(**b**) Metaplots showing strong enrichment of retrotransposable elements (RTEs) in sense orientation immediately upstream of regulated iCSs (red line ±SEM) and their relative depletion in the CS-proximal region on the antisense strand (blue line ±SEM). Note that the antisense RTE density values were multiplied by -1.

(**c**) iCS-associated RTEs (sense-strand RTEs terminating in ±50-nt vicinity of regulated iCSs) are enriched for SINEs as compared to the overall incidence of these elements in regulated first introns or the entire genome. (a-c) iCSs were considered regulated if they were upregulated in response to siSrrt ≥2-fold, FDR<0.05 and their host gene was downregulated ≥1.5-fold, FDR<0.05.

(**d, e**) Examples of Srrt-dependent genes with iCSs matching 3' ends of sense-strand (d) SINEs or (e) LINEs. RNA-Seq coverage plots are shown in gray and 3'RNA-Seq data are in red. Similar to the genes with conserved iCSs in Fig. 2d, upregulation of RTE-associated iCSs in response to siSrrt leads to a pronounced decrease in the RNA-Seq and 3'RNA-Seq signals in corresponding 3'UTRs. Red arrowheads, CSs preceded by AATAAA or ATTAAA hexamers. Pairwise alignments between regulated CSs and corresponding RTE consensus sequences are shown at the bottom of each panel with invariant positions marked by vertical bars and degenerate matches and base transitions indicated by colons. Canonical PAS hexamers are highlighted in pink.

Figure 8. Recurrent retrotransposition may increase gene dependence on Srrt

(**a**) Regulated iCSs associated with 3' ends of sense-strand RTEs show significantly lower evolutionary conservation (PhastCons) score than other regulated iCSs.

(**b**) Sense-strand RTEs terminated in iCS vicinity are typically less divergent from the corresponding master copies than control groups.

(**c**) The overall RTE density is significantly higher in Srrt-regulated first introns than in non-regulated first or non-first introns. Also note a strong bias towards antisense orientation of RTEs in all groups of introns.

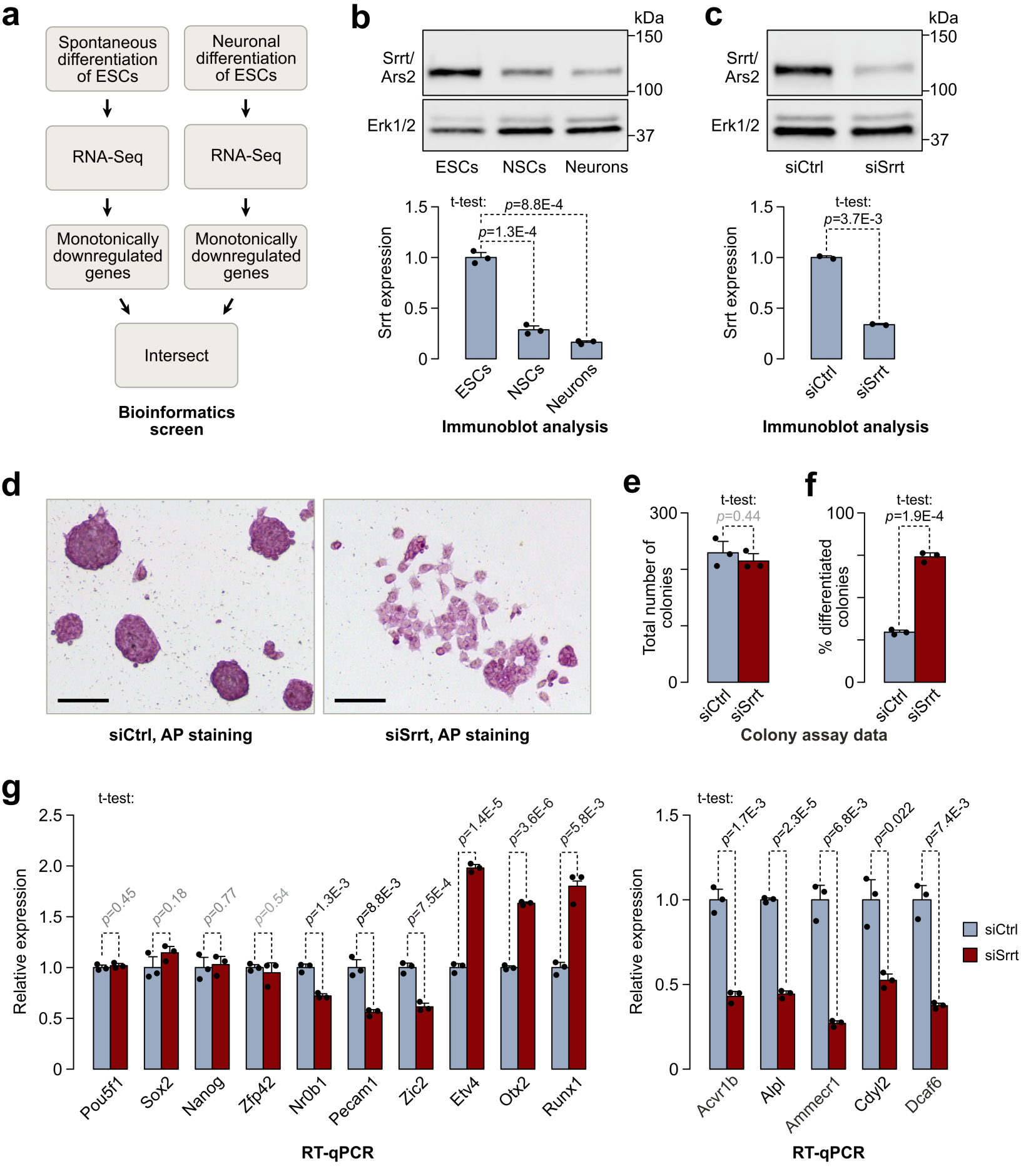
(**d**) Length of first introns positively correlates with the percent of sequence occupied by RTEs on both strands. Dashed line, linear regression.

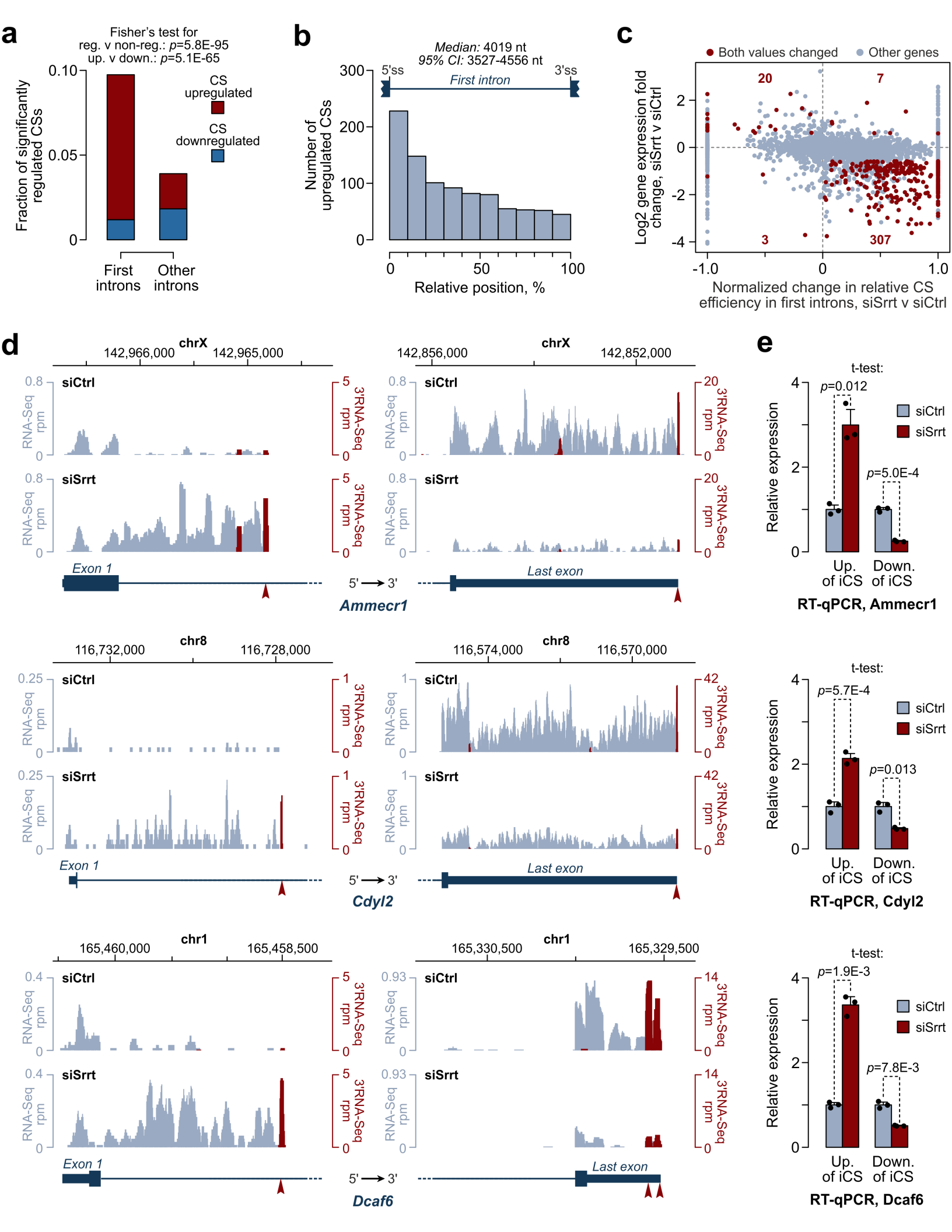
(**e**) Consistent with their higher RTE load, the length of first introns in Srrt-dependent genes tends to exceed that of non-regulated or non-first introns. (a-c, e) iCSs were considered regulated if they were upregulated in response to siSrrt ≥2-fold, FDR<0.05 and their host gene was downregulated ≥1.5-fold, FDR<0.05. In (a-c) and (e), box bounds, the first and the third quartiles; thick black lines, the medians. Whiskers extend from the first and the third quartile to the lowest and highest data points or, if there are outliers, 1.5× of the interquartile range. Outliers are not shown.

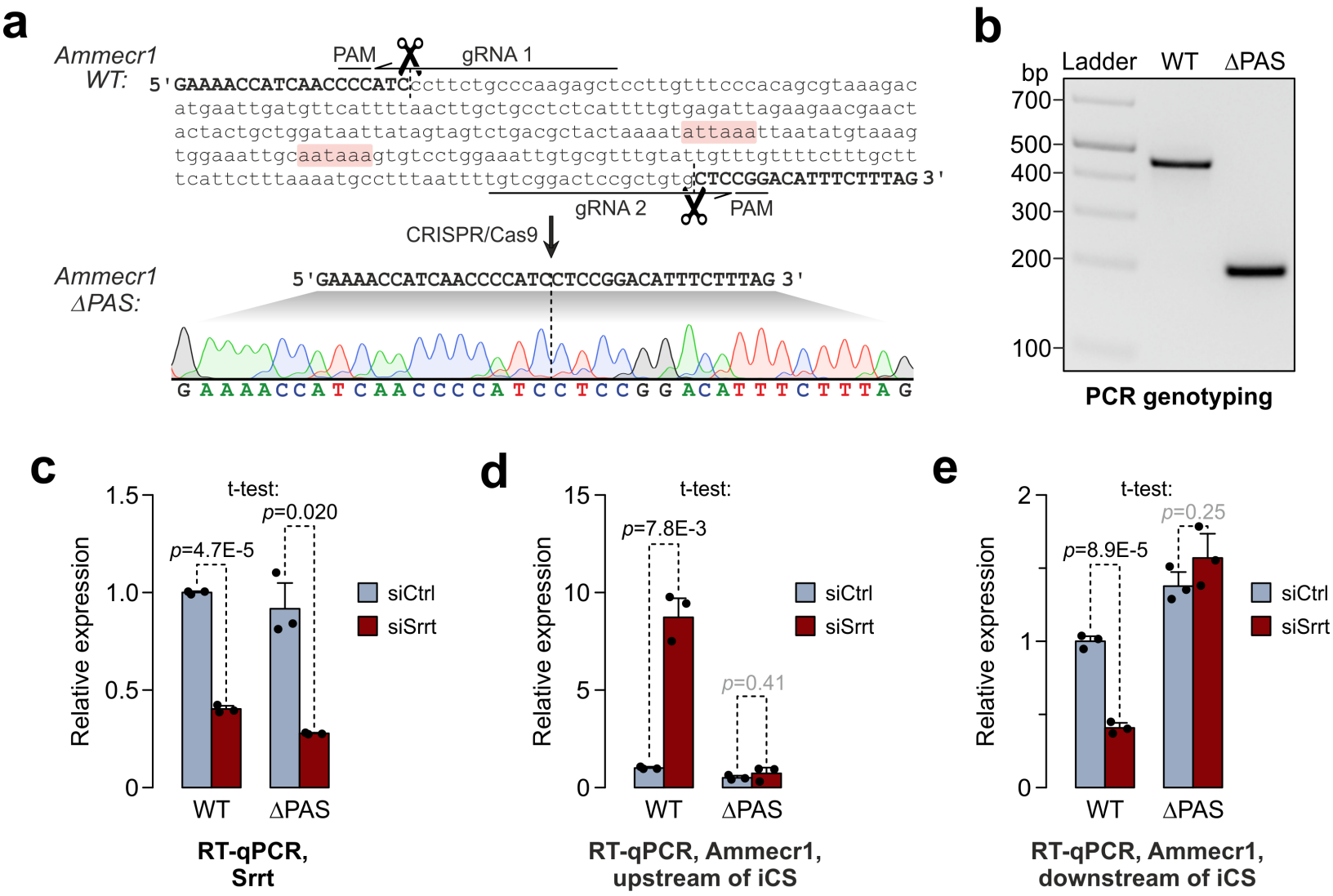
(**f**) Gene expression in ESCs shows a negative relationship with the length of the first intron even in the presence of normal amounts of Srrt. Shown are mean expression values ±SEM in siCtrl-treated ESCs for genes with short (shorter than the 1/3 quantile; i.e. <1524 nt), midsize (i.e. longer than or equal to the 1/3 quantile but shorter than the 2/3 quantile; i.e. ≥1524 and <7251 nt), and long first introns (longer than or equal to the 2/3 quantile; i.e. ≥7251 nt). Note that genes with AATAAA(s) in the first intron are expressed at levels statistically indistinguishable from their AATAAA-free counterparts.

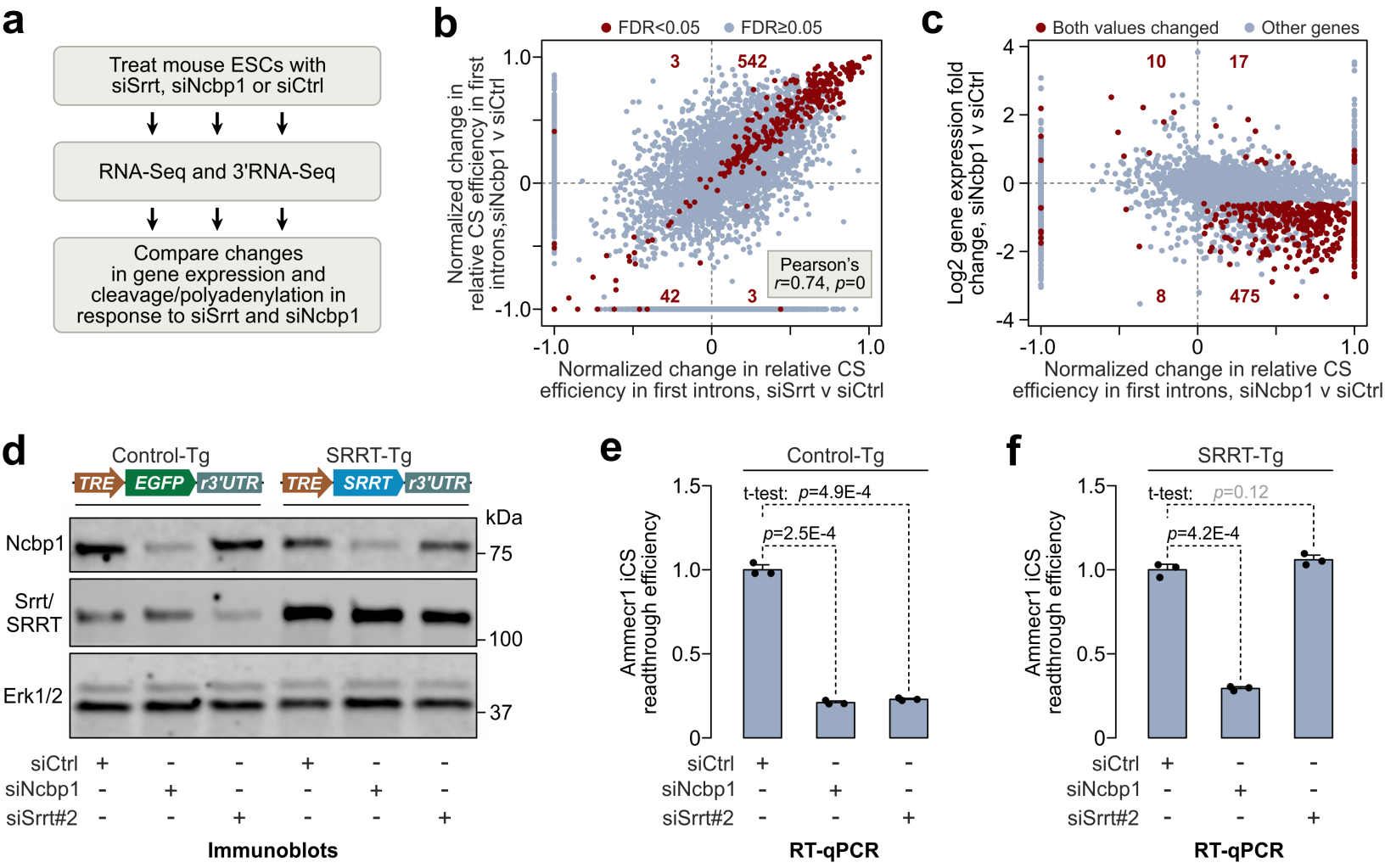
(**g**) Srrt knockdown leads to preferential downregulation of genes with long first introns containing at least one AATAAA hexamer.

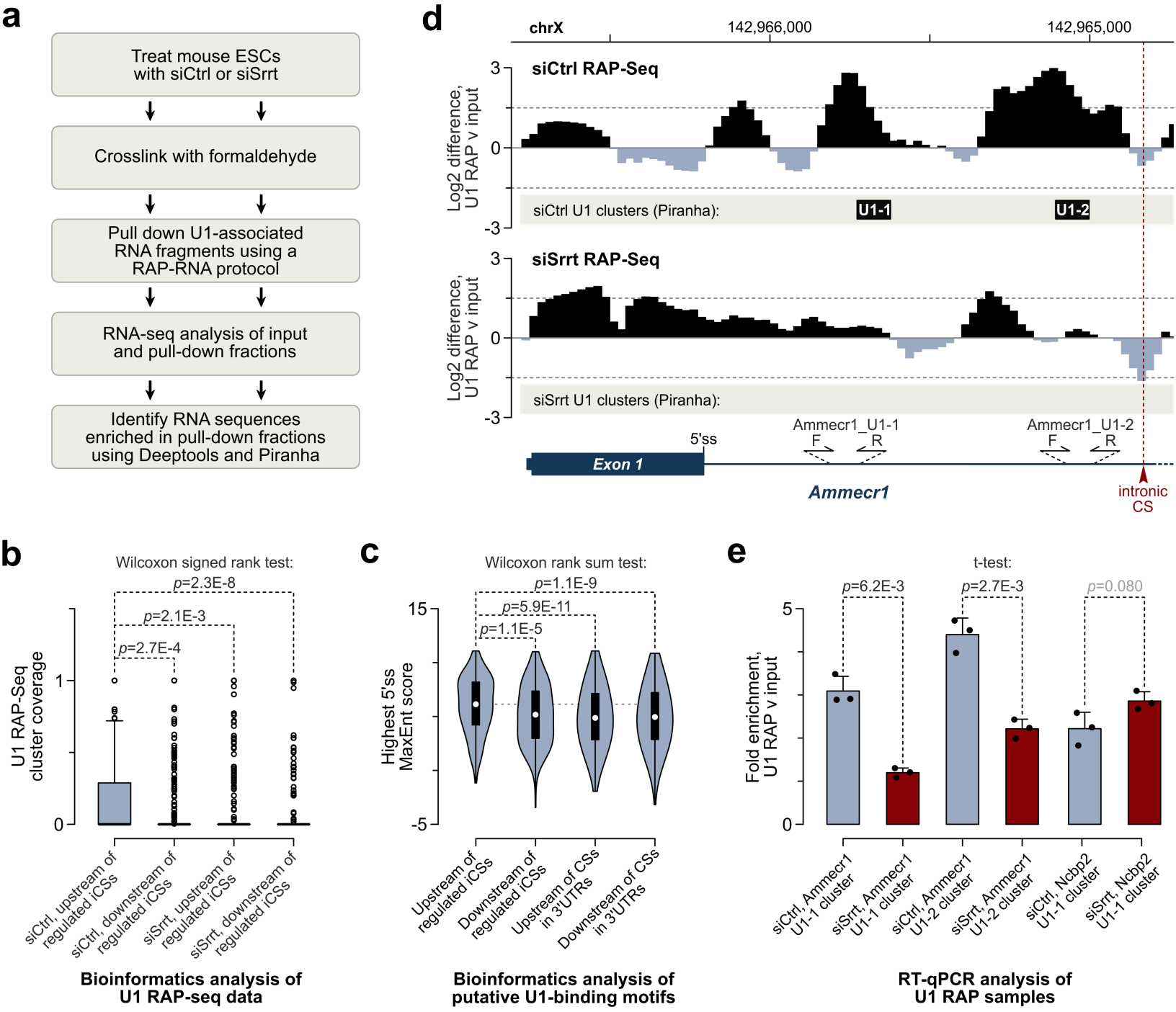
(**h**) Naturally high levels of Srrt help ESCs to maintain their gene expression program through a transcription antitermination mechanism.

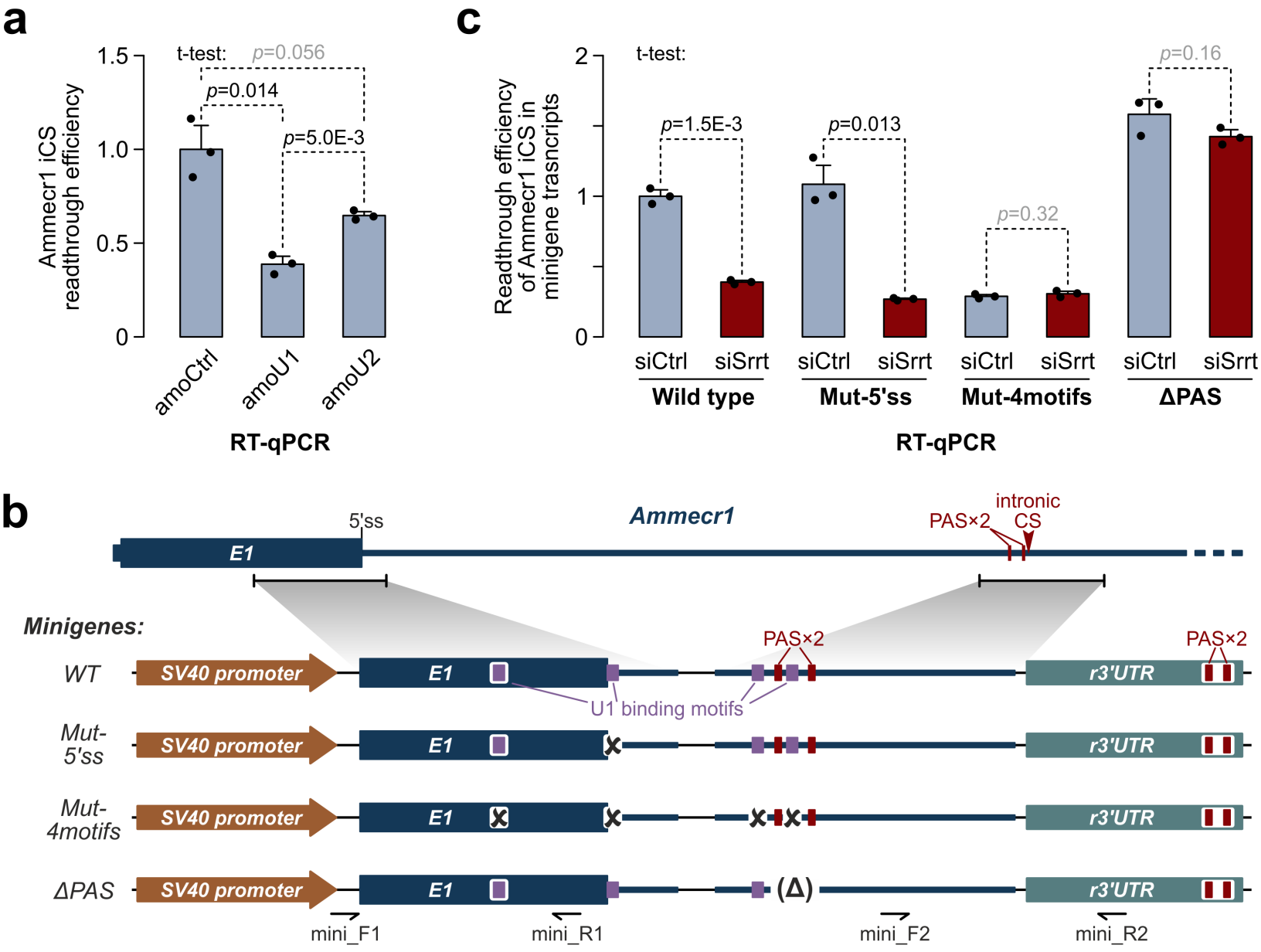

a
Spontaneous differentiation of ESCs
Neuronal differentiation of ESCs
RNA-Seq
RNA-Seq
Monotonically downregulated genes
Monotonically downregulated genes
Intersect
Bioinformatics screen
b
kDa
150
100
37
Srrt/ Ars2
Erk1/2
ESCs
NSCs
Neurons
t-test:
p=8.8E-4
p=1.3E-4
Srrt expression
1.5
1.0
0.5
0
ESCs
NSCs
Neurons
Immunoblot analysis
c
kDa
150
100
37
Srrt/ Ars2
Erk1/2
siCtrl
siSrrt
t-test:
p=3.7E-3
Srrt expression
1.5
1.0
0.5
0
siCtrl
siSrrt
Immunoblot analysis
d
siCtrl, AP staining
siSrrt, AP staining
e
t-test:
p=0.44
Total number of colonies
300
0
siCtrl
siSrrt
f
t-test:
p=1.9E-4
% differentiated colonies
100
0
siCtrl
siSrrt
Colony assay data
g
t-test:
p=0.45
p=0.18
p=0.77
p=0.54
p=1.3E-3
p=8.8E-3
p=7.5E-4
p=1.4E-5
p=3.6E-6
p=5.8E-3
Relative expression
2.5
2.0
1.5
1.0
0.5
0
Pou5f1
Sox2
Nanog
Zfp42
Nr0b1
Pecam1
Zic2
Etv4
Otx2
Runx1
RT-qPCR
t-test:
p=1.7E-3
p=2.3E-5
p=6.8E-3
p=0.022
p=7.4E-3
Relative expression
1.5
1.0
0.5
0
Acvr1b
Alpl
Ammecr1
Cdyl2
Dcaf6
RT-qPCR
siCtrl
siSrrt

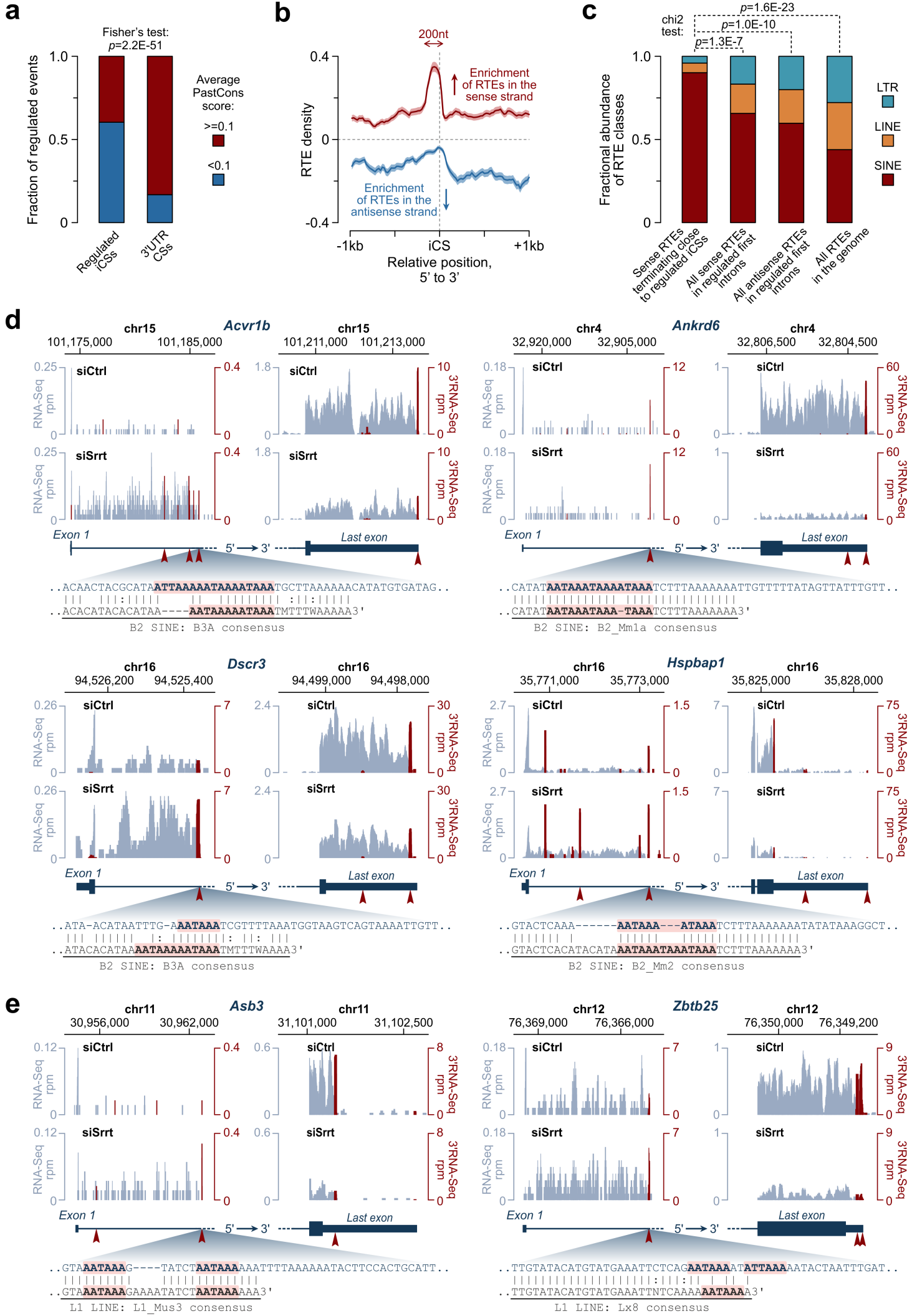

a
Fisher's test for
reg. v non-reg.: p=5.8E-95
up. v down.: p=5.1E-65
Fraction of significantly regulated CSs
CS upregulated
CS downregulated
First introns
Other introns
b
Median: 4019 nt
95% CI: 3527-4556 nt
5'ss
3'ss
First intron
Number of upregulated CSs
Relative position, %
c
Both values changed
Other genes
20
7
3
307
Log2 gene expression fold change, siSrrt v siCtrl
Normalized change in relative CS efficiency in first introns, siSrrt v siCtrl
d
chrX
142,966,000
142,965,000
142,856,000
142,852,000
siCtrl
siSrrt
RNA-Seq rpm
3'RNA-Seq rpm
Exon 1
Last exon
5' → 3'
Ammecr1
chr8
116,732,000
116,728,000
116,574,000
116,570,000
Cdyl2
chr1
165,460,000
165,458,500
165,330,500
165,329,500
Dcaf6
e
t-test:
p=0.012
p=5.0E-4
Relative expression
siCtrl
siSrrt
Up. of iCS
Down. of iCS
RT-qPCR, Ammecr1
p=5.7E-4
p=0.013
RT-qPCR, Cdyl2
p=1.9E-3
p=7.8E-3
RT-qPCR, Dcaf6

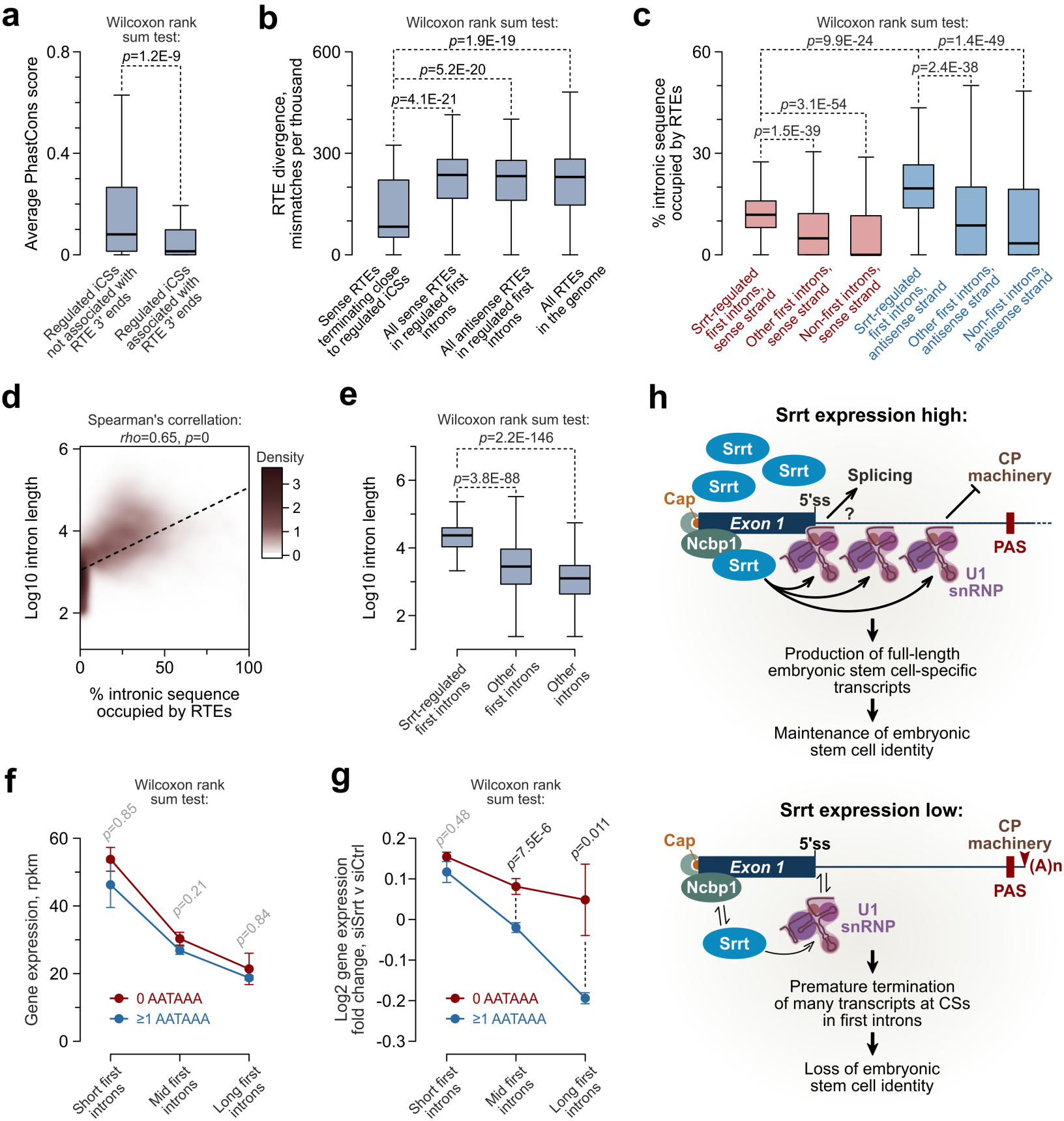

a
Ammecr1 WT:
PAM
gRNA 1
5' GAAAACCATCAACCCCATCcttctgcccaagagctccttgtttcccacagcgtaaagac
atgaattgatgttcatttaacttgctgcctctcattttgtgagattagaagaacgaact
actactgctggataattatagtagtctgacgctactaaaatattaaattaatatgtaaag
tggaaattgcaataaagtgtcctggaaattgtgcgtttgtattgtttgttttctttgctt
tcattctttaaaatgcctttaattttgtcggactccgctgtgCTCCGGACATTTCTTTAG 3'
gRNA 2
PAM
CRISPR/Cas9
Ammecr1 ΔPAS:
5' GAAAACCATCAACCCCATCCTCCGGACATTTCTTTAG 3'
G A A A A C C A T C A A C C C C A T C C T C C G G A C A T T T C T T T A G
b
bp
Ladder
WT
ΔPAS
700
500
400
300
200
100
PCR genotyping
c
t-test:
p=4.7E-5
p=0.020
Relative expression
1.5
1.0
0.5
0
siCtrl
siSrrt
WT
ΔPAS
RT-qPCR, Srrt
d
t-test:
p=7.8E-3
p=0.41
Relative expression
15
10
5
0
siCtrl
siSrrt
WT
ΔPAS
RT-qPCR, Ammecr1, upstream of iCS
e
t-test:
p=8.9E-5
p=0.25
Relative expression
2
1
0
siCtrl
siSrrt
WT
ΔPAS
RT-qPCR, Ammecr1, downstream of iCS


a
Treat mouse ESCs with siSrrt, siNcbp1 or siCtrl
RNA-Seq and 3'RNA-Seq
Compare changes in gene expression and cleavage/polyadenylation in response to siSrrt and siNcbp1
b
FDR<0.05
FDR≥0.05
Normalized change in relative CS efficiency in first introns,siNcbp1 v siCtrl
Normalized change in relative CS efficiency in first introns, siSrrt v siCtrl
Pearson's r=0.74, p=0
3
542
42
3
c
Both values changed
Other genes
Log2 gene expression fold change, siNcbp1 v siCtrl
Normalized change in relative CS efficiency in first introns, siNcbp1 v siCtrl
10
17
8
475
d
Control-Tg
SRRT-Tg
TRE
EGFP
r3'UTR
SRRT
kDa
Ncbp1
Srrt/ SRRT
Erk1/2
75
100
37
siCtrl
siNcbp1
siSrrt#2
Immunoblots
e
Control-Tg
t-test: p=4.9E-4
p=2.5E-4
Ammecr1 iCS readthrough efficiency
RT-qPCR
f
SRRT-Tg
t-test: p=0.12
p=4.2E-4
Ammecr1 iCS readthrough efficiency
RT-qPCR


a
Treat mouse ESCs with siCtrl or siSrrt
Crosslink with formaldehyde
Pull down U1-associated RNA fragments using a RAP-RNA protocol
RNA-seq analysis of input and pull-down fractions
Identify RNA sequences enriched in pull-down fractions using Deeptools and Piranha
d
chrX
142,966,000
142,965,000
siCtrl RAP-Seq
Log2 difference, U1 RAP v input
siCtrl U1 clusters (Piranha):
U1-1
U1-2
siSrrt RAP-Seq
siSrrt U1 clusters (Piranha):
5'ss
Ammecr1_U1-1
Ammecr1_U1-2
F
R
Exon 1
Ammecr1
intronic CS
b
Wilcoxon signed rank test:
p=2.3E-8
p=2.1E-3
p=2.7E-4
U1 RAP-Seq cluster coverage
siCtrl, upstream of regulated iCSs
siCtrl, downstream of regulated iCSs
siSrrt, upstream of regulated iCSs
siSrrt, downstream of regulated iCSs
Bioinformatics analysis of U1 RAP-seq data
c
Wilcoxon rank sum test:
p=1.1E-9
p=5.9E-11
p=1.1E-5
Highest 5'ss MaxEnt score
Upstream of regulated iCSs
Downstream of regulated iCSs
Upstream of CSs in 3'UTRs
Downstream of CSs in 3'UTRs
Bioinformatics analysis of putative U1-binding motifs
e
t-test:
p=6.2E-3
p=2.7E-3
p=0.080
Fold enrichment, U1 RAP v input
siCtrl, Ammecr1 U1-1 cluster
siSrrt, Ammecr1 U1-1 cluster
siCtrl, Ammecr1 U1-2 cluster
siSrrt, Ammecr1 U1-2 cluster
siCtrl, Ncbp2 U1-1 cluster
siSrrt, Ncbp2 U1-1 cluster
RT-qPCR analysis of U1 RAP samples


a
Ammecr1 iCS readthrough efficiency
1.5
1.0
0.5
0
t-test:
p=0.056
p=0.014
p=5.0E-3
amoCtrl
amoU1
amoU2
RT-qPCR
c
Readthrough efficiency of Ammecr1 iCS in minigene trasncripts
2
1
0
t-test:
p=1.5E-3
p=0.013
p=0.32
p=0.16
siCtrl
siSrrt
Wild type
Mut-5'ss
Mut-4motifs
ΔPAS
RT-qPCR
b
Ammecr1
5'ss
intronic CS
PAS×2
E1
Minigenes:
WT
Mut-5'ss
Mut-4motifs
ΔPAS
SV40 promoter
r3'UTR
U1 binding motifs
(Δ)
mini_F1
mini_R1
mini_F2
mini_R2


a
Fisher's test:
p=2.2E-51
Fraction of regulated events
1.0
0.5
0
Average PastCons score:
>=0.1
<0.1
Regulated iCSs
3'UTR CSs
b
200nt
0.4
0
-0.4
RTE density
Enrichment of RTEs in the sense strand
Enrichment of RTEs in the antisense strand
-1kb
iCS
+1kb
Relative position, 5' to 3'
c
chi2 test:
p=1.6E-23
p=1.0E-10
p=1.3E-7
Fractional abundance of RTE classes
1.0
0.5
0
LTR
LINE
SINE
Sense RTEs terminating close to regulated iCSs
All sense RTEs in regulated first introns
All antisense RTEs in regulated first introns
All RTEs in the genome
d
Acvr1b
chr15
101,175,000
101,185,000
chr15
101,211,000
101,213,000
siCtrl
siSrrt
RNA-Seq rpm
3'RNA-Seq rpm
Exon 1
Last exon
5'
3'
..ACAACTACGCATAATTAAAAATAAAATAAATGCTTAAAAAACATATGTGATAG..
..ACACATACACATAA----AATAAAAATAAATMTTTWAAAAA3'
B2 SINE: B3A consensus
Ankrd6
chr4
32,920,000
32,905,000
chr4
32,806,500
32,804,500
..CATATAATAAATAAAATAAATCTTTAAAAAAAATTGTTTTTATAGTTATTTGTT..
..CATATAATAAATAAA-TAAATCTTTAAAAAAA3'
B2 SINE: B2_Mm1a consensus
Dscr3
chr16
94,526,200
94,525,400
chr16
94,499,000
94,498,000
..ATA-ACATAATTTG-AAATAAATCGTTTTAAATGGTAAGTCAGTAAAATTGTT..
..ATACACATAAAATAAAAATAAATMTTTWAAAA3'
B2 SINE: B3A consensus
Hspbap1
chr16
35,771,000
35,773,000
chr16
35,825,000
35,828,000
..GTACTCAAA------AATAAA---ATAAATCTTTAAAAAAATATATAAAGGCT..
..GTACTCACATACATAAATAAATAAATAAATCTTTAAAAAAA3'
B2 SINE: B2_Mm2 consensus
e
Asb3
chr11
30,956,000
30,962,000
chr11
31,101,000
31,102,500
..GTAAATAAAG----TATCTAATAAAAAATTTTAAAAAATACTTCCACTGCATT..
..GTAAATAAAGAAAATATCTAATAAAAAA3'
L1 LINE: L1_Mus3 consensus
Zbtb25
chr12
76,369,000
76,366,000
chr12
76,350,000
76,349,200
..TTGTATACATGTATGAAATTCTCAGAATAAAATATTTAAAAATACTAATTTGAT..
..TTGTATACATGTATGAAATTNTCAAAAAATAAAA3'
L1 LINE: Lx8 consensus


a
Wilcoxon rank sum test:
p=1.2E-9
Average PhastCons score
0.8
0.4
0
Regulated iCSs not associated with RTE 3' ends
Regulated iCSs associated with RTE 3' ends
b
Wilcoxon rank sum test:
p=1.9E-19
p=5.2E-20
p=4.1E-21
RTE divergence, mismatches per thousand
600
300
0
Sense RTEs terminating close to regulated iCSs
All sense RTEs in regulated first introns
All antisense RTEs in regulated first introns
All RTEs in the genome
c
Wilcoxon rank sum test:
p=9.9E-24
p=1.4E-49
p=2.4E-38
p=3.1E-54
p=1.5E-39
% intronic sequence occupied by RTEs
60
30
0
Srrt-regulated first introns, sense strand
Other first introns, sense strand
Non-first introns, sense strand
Srrt-regulated first introns, antisense strand
Other first introns, antisense strand
Non-first introns, antisense strand
d
Spearman's correllation:
rho=0.65, p=0
Density
3
2
1
0
Log10 intron length
6
4
2
0
50
100
% intronic sequence occupied by RTEs
e
Wilcoxon rank sum test:
p=2.2E-146
p=3.8E-88
Log10 intron length
6
4
2
Srrt-regulated first introns
Other first introns
Other introns
f
Wilcoxon rank sum test:
p=0.85
p=0.21
p=0.84
Gene expression, rpkm
60
40
20
0
0 AATAAA
≥1 AATAAA
Short first introns
Mid first introns
Long first introns
g
Wilcoxon rank sum test:
p=0.48
p=7.5E-6
p=0.011
Log2 gene expression fold change, siSrrt v siCtrl
0.2
0.1
0
-0.1
-0.2
-0.3
0 AATAAA
≥1 AATAAA
Short first introns
Mid first introns
Long first introns
h
Srrt expression high:
Srrt
Srrt
Srrt
Splicing
CP machinery
Cap
5'ss
Exon 1
?
Ncbp1
Srrt
PAS
U1 snRNP
Production of full-length embryonic stem cell-specific transcripts
Maintenance of embryonic stem cell identity
Srrt expression low:
Cap
5'ss
CP machinery
Exon 1
(A)n
Ncbp1
PAS
U1 snRNP
Srrt
Premature termination of many transcripts at CSs in first introns
Loss of embryonic stem cell identity